We might be wrong, but we think that hedging doesn't protect your reputation

Peter J. Collins[1]

Ulrike Hahn1[1,2]

[1]Munich Center for Mathematical Philosophy

[2]Birkbeck, University of London

Author Note

Peter Collins is now at the Department of Psychology, Goldsmiths, University of London

Corresponding Author:

Dr Peter Collins

Department of Psychology

Goldsmiths, University of London

New Cross

London

SE14 6NW

p.j.collins.65@gmail.com

Copies of the surveys, data files and analysis scripts can be found on the Open Science Framework at osf.io/r9cna.

Abstract

We gain much of our knowledge from other people. Since people are fallible - they lie, mislead, and are mistaken - it seems essential to monitor their claims and their reliability as sources of information. An intuitive way to do this is to draw on our expectations about claims and sources: to perform expectation-based updating (Hahn, Merdes, & von Sydow, 2018). But this updating can have damaging consequences, leading us into a kind of confirmation bias. An alternative is to keep track of outcomes and record whether a claim proves true or false: to perform outcome-based updating (Hahn et al., 2018). This form of updating does not have the negative repercussions on belief accuracy. But both forms of updating might undermine the trust and cooperation assumed to be necessary for successful communication. We explore a potential boundary condition on these types of updating. We investigate whether speakers can protect their reputation when they make claims with low prior probability, with and without knowledge of the final outcome. We explore suggestions from McCready (2015) that speakers can protect themselves by hedging with evidential language: in particular with weaker propositional attitudes ('I suspect that...') and so-called double hedges ('I might be wrong, but I think...'). We find that both forms of updating are robust to hedging with this evidential language, and find no clear evidence for a protective effect. We discuss extra ingredients that may allow successful hedging.

Keywords: cooperation; hedging; pragmatics; testimony; belief change

Each day we receive information from other people, from the mundane ("There is heavy traffic on the motorway") to the highly consequential ("This pension scheme will give you security in retirement"). We can use this information to form or change our beliefs. When we update our beliefs in this way, we are engaging with people's testimony: with their saying, telling, or asserting; their committing to the truth of some proposition intending that a recipient (or recipients) will accept the proposition as true (see, e.g., Adler, 2015; Coady, 1992). Testimony seems both ubiquitous and essential[1]. Most of us have not experienced that the world is spherical, witnessed the events in the news, collected evidence of historical events, or studied the safety of medical treatments. But much of the time, we will accept other people's testimony on such matters. Testimony allows a division of labour: we all want to know how to treat our ailments, but we cannot all be doctors. But it also raises a normative question: when should we trust, and when distrust, our information sources?

Information sources can clearly be unreliable, leaving us at risk of being misinformed, misled, or deceived. Sources can, for instance, be well-intentioned but inaccurate: people forget, and their memories are reconstructive (for discussion, see Hahn, Oaksford, & Harris, 2012; on reconstructive memory, see, e.g., Loftus, 1975). Sources can also be dishonest: people routinely engage in small acts of deception, cheating when they can do so without being caught or without challenging their self-

[1] This section draws substantially on (Collins, 2017).

conception (see, e.g., Chance, Norton, Gino, & Ariely, 2011; Mazar, Amir, & Ariely, 2008).

Systematic unreliability seems to compel us to be "epistemically vigilant": to monitor the source and content of messages, and to distinguish between comprehending and accepting a message (Sperber et al., 2010). Although this vigilance may be fragile (on this debate, see Gilbert, Krull, & Malone, 1990; Gilbert, Tafarodi, & Malone, 1993; Hasson, Simmons, & Todorov, 2005; Mandelbaum, 2014), we clearly can be epistemically vigilant. For instance, developmental studies show that even 14-month-olds can distinguish between reliable and unreliable informants (Poulin-Dubois, Brooker, & Polonia, 2011; Poulin-Dubois & Chow, 2009); that 3-year-olds are sensitive to verbal uncertainty (Sabbagh & Baldwin, 2001), expertise (Robinson, Champion, & Mitchell, 1999), and accuracy (Ganea, Koenig, & Millett, 2011); and that older children (between 6 and 8 years' old) are sensitive to deception (Mascaro & Sperber, 2009), self-interest (Mills & Keil, 2005), and partiality (Mills & Keil, 2008). Such vigilance continues into adulthood, as shown in experimental studies of persuasion (for recent reviews, see Briñol & Petty, 2009; Petty & Briñol, 2008).

One form of vigilance is to monitor a source's claims and the actual outcomes: whether the claims prove true or false. The recipient can use the outcomes to revise their judgments about a source's reliability. When a source's claim (hypothesis, belief) proves to be true, the source's reliability increases; when a source's claim proves to be false, the source's reliability decreases. Following Hahn, Merdes, and von Sydow (2018) we call this form of vigilance "outcome-based updating". Outcome-based updating is highly effective. So long as the recipient monitors claims and eventual

outcomes, the recipient will converge on a true belief about the reliability of the source (Hahn, Merdes, & von Sydow, 2018). But outcome-based updating does not apply when outcomes are unknown or when reasoning about a singular event, such as an individual piece of witness testimony in court (Hahn et al., 2018).

Epistemic vigilance is also embodied in a form of updating seen in putatively normative models of testimony from social epistemology: the models of Bovens and Hartmann (2003) and Olsson and Angere (reported in Olsson, 2011; Olsson & Vallinder, 2013). These models capture belief change from a source's assertion in terms of subjective probability (degrees of belief). The models both assume that source reliability and information content interact. In these models, the recipient of some claim (proposition, assertion, statement) assigns subjective probabilities to two key variables: the reliability of the source and the information content of the claim. Source reliability is understood as the (subjective) probability that a particular source will say something true. The recipient has (or can form) an initial judgment of reliability, prior to the specific claim being made; the recipient can, further, revise that judgment in light of the claim. With respect to the information content, the recipient has (or can form) some subjective prior probability. We can gloss this prior as the degree of belief in the claim before receiving any evidence, whether from observation or someone's assertion; or as how expected it is that the claim will prove true, again before receiving any such evidence. We will use the terms "high expectedness" and "low expectedness"[2] to indicate the degree to which the claim is expected to prove true. The recipient updates the prior in light of the claim and the source's reliability.

[2] In this paper we deal exclusively with binary hypotheses, for which "low prior" (low expectedness) and "high prior" (high expectedness) refer to either side of the midpoint of the

On these models, reliable sources should tend to increase belief[3]; high prior (high expectedness) information should tend to increase perceived reliability; and low-prior (low expectendness) information should tend to decrease perceived reliability. Following Hahn et al. (2018) we call this form of vigilance "expectation-based updating".

Expectation-based updating can apply with single events, such as witness testimony in a courtroom, where a single underlying historic event is at issue. This updating only requires that a recipient has - or can generate - (subjective) prior probabilities for the source's reliability and for the claim in hand. Since these probabilities are subjective, they do not need to refer to relative frequencies of events. However, expectation-based updating can have problematic effects. Through time, sources may display a kind of confirmation bias, amplifying (subjectively) plausible evidence and down-weighting (subjectively) implausible evidence (Hahn et al., 2018).

Given that people seem to perform expectation-based updating (see, e.g., Collins, Hahn, von Gerber, & Olsson, 2018), perhaps we should worry about testimony: perhaps our natural, intuitively rational tendencies lead us astray. But expectation-based updating has been evidenced so far with claims that are bold assertions: assertions that admit no possibility of error (Collins et al., 2018). Claims do not, of course, have to be so strong. Their sources can draw on language for

---

scale (.5). The midpoint reflects agnosticism: the hypothesis is as likely to be true as false. For multivalued hypotheses - for instance, that a person has a specific telephone number - the relevant midpoint separating low/high prior (expectedness) is 1/N, where N is the number of rival hypotheses considered.

[3] The models differ in their approach to unreliable sources: the Bovens and Hartmann model assumes that hearers, in effect, randomize on whether to believe a maximally unreliable source; the Olsson and Angere model assumes that hearers will decrease their belief in information from a maximally unreliable source.

making weaker, less certain claims. Perhaps, then, such language establishes boundary conditions on expectation-based updating. We will consider one boundary condition in the experiments reported below.

Expectation-based updating and outcome-based updating share a potential disadvantage for speakers, in that they may undermine the basis of successful communication. It is widely assumed that communication is, in some sense, cooperative (Grice, 1975; McCready, 2015). This assumption can be fleshed out in various ways. As Sperber et al. (2010) put it, speakers and hearers cooperate by investing effort: speakers invest effort to communicate something, and hearers invest effort to attend to and interpret it. This cooperative effort is sustained by the expectation of benefit for each party. The benefit for the speaker is to produce some desired effect; the benefit for the hearer is to acquire true, relevant information (Sperber et al., 2010).

Fulfilling the hearer's expectations of benefit is no mean feat. In particular, truth and relevance can be conflicting goals. Imagine a speaker who prioritizes truthfulness. That speaker might decide only to communicate information they are certain about. When that speaker has information that is relevant but uncertain, they would withhold that information. If the withholding comes to light, would the speaker not suffer reputational damage? There is no guarantee that speakers possess information that is both relevant and certain. But we expect speakers to be informative, and an uninformative speaker may damage their reputation (McCready, 2015). Conversely, if a speaker decides to communicate relevant but uncertain information, they may also damage their reputation if that information turns out to be

false (McCready, 2015) or merely has a low prior probability (Collins, Hahn, von Gerber, & Olsson, 2015; Collins et al., 2018).

Cooperation is integral to recent formal models of communication: models that use the tools of game theory (e.g. Goodman & Frank, 2016; McCready, 2015). One such model is the Rational Speech Act Model (RSA), which has had considerable success in explaining phenomena such as scalar implicatures, hyperbole, and vagueness (for discussion, see, e.g., Goodman & Frank, 2016). This model is premised on cooperation. Simplifying somewhat, hearer and speaker reason about each other[4], and hearers assume that the speaker chooses an utterance proportionately to its expected utility, which is determined by the "social benefit of providing epistemic help to the listener" (Goodman & Frank, 2016, p. 820).

But what happens to cooperation when a hearer downgrades their belief in a speaker's reputation? McCready (2015) points out a fundamental dilemma raised by game-theoretic accounts. Cooperation requires speakers to communicate relevant information, but when that information turns out to be false, it threatens the mere possibility of subsequent communication. When a speaker's reputation drops enough, that speaker may struggle to establish enough cooperation for communication to proceed (McCready, 2015). A speaker with a bad reputation will presumably seem unlikely to offer truthful, relevant information[5] - information that motivates hearers to attend and interpret (Sperber et al., 2010). So speakers must balance two, often conflicting goals, cooperation and reputation management. This means that speakers

[4] In fact, in this model the speaker reasons about an imagined literal hearer to avoid an infinite regress.

[5] We will focus in this paper on truth, rather than relevance, and on hedging of truth. See McCready (2015) for illuminating discussion of hedging beyond truth.

must possess a strategy for trying to "ensure that honest mistakes don't destroy the possibility of future trust and cooperation"(McCready, 2015, p.3) - and for "signal[ing] that they do not wish to take responsibility for the signals they use" (McCready, 2015, p.39). One such strategy, McCready notes, is the use of *hedges*: evidential language ("I suspect X is the case...") that may serve as "grammatical mechanisms for protecting reputations" (McCready, 2015, p. 3).

Hedging

The term "hedging" has a number of difference senses. Hedging is said to occur when speakers qualify their assertions to lessen the impact in some way: for instance, to soften bad news, to make an utterance more polite, or to avoid a firm commitment to the truth or to an action (see, e.g., Holtgraves, 2002). We restrict our attention, here, to hedging which qualifies assertions to allow for exceptions or possible falsehood (McCready, 2015). This hedging can be likened to disclaimers in advertising (McCready, 2015). Take, for instance, the following examples (the hedges are italicized):

(a) John is *sort of* stupid

(b) *I suspect that* it is cold outside

(c) *I might be wrong, but* Palin is not going to be elected.

(d) *This might not be true, but* she doesn't really care about you.

(McCready, 2015, p. 39)

McCready (2015) argues that hedging with evidential language allows speakers to maintain cooperation without having to be perfectly reliable sources. Hedging does so by allowing sources to protect their reputations. Reputations and reliability, in McCready's theory, correspond well to the notion of reliability seen in

the Bayesian models of testimony above. McCready takes hearers to remember their interactions with particular sources; to record whether sources' claims proved true, false, or indeterminate; and to judge the proportion of claims that proved true. This proportion is then the reliability of the source (McCready, 2015, p. 9). Reliability, in this sense, can easily be understood as the prior probability, or base rate, that a particular source will say something true. When hearers interact with a new source, they have no history of interactions to draw on; but they can presumably use background information, such as stereotypes[6], to form a rapid judgment of reliability.

A wide range of evidential language could, in principle, function as hedges, but we focus on two types of evidential hedge distinguished by McCready (2015) which can be seen in examples (a) to (d) above. On McCready's account, while all the hedges act as disclaimers, hedges (a) and (b) work differently from (c) and (d). Hedges (a) and (b), McCready argues, modify the assertions directly: they soften the embedded proposition to the point at which it is assertable. Hedges (c) and (d), in contrast, modify the assertions more indirectly: a bold assertion is made, but a 'shield hedge', or disclaimer, is added (McCready, 2015). Hedges (c) and (d) do not, however, strike all hearers as acceptable (McCready, 2015). For such hearers - the present authors included - a second hedge is needed. Hence:

(e) *I might be wrong, but I think* that Palin is not going to be elected.

(f) *This might not be true, but I think* that she doesn't really care about you.

McCready terms these 'double hedges'. While McCready (2015) focuses on shielding,

[6] We are not implying that stereotypes provide accurate information (though see, e.g., Jussim, Crawford, & Rubinstein, 2015, for such a claim), but they nevertheless provide information that hearers might presume is accurate and use accordingly.

we consider both types of evidential hedge, since both offer a way to defuse the reputational hit and its potential damage to cooperation.

What, we might ask, is the semantic content of these hedges? We will assume that these hedges express considerably less than full certainty, which we will understand as a probability considerably below 1. It is a widespread assumption in psychology that expressions of uncertainty can be modelled with probabilities, as shown in the large literature on verbal probability expressions ("It is virtually impossible/possible/likely/virtually certain")(see, e.g., Budescu & Wallsten, 1985; Karelitz & Budescu, 2004; Wallsten & Budescu, 1995). Similarly, probabilities feature in a recent account of epistemic modality (Lassiter, 2010, 2017). We do not have conceptual grounds for fixing an exact point value for this probability. Indeed, we take research on verbal probabilities to show how difficult it is to associate such words with a fixed point value or range. But a loose point estimate of (somewhere around) a "60% chance" seems reasonable[7].

This account of hedging as reputation management is intuitively appealing and compatible with existing research. Theories of argumentation, for instance, hold that it is a key argumentative skill to calibrate claims to the strength of evidence or certainty using appropriate evidential language (Kuhn, 1991; Toulmin, Rieke, & Janik, 1979). It seems to follow that a speaker using such language should be given credit for doing so and should not be treated as though they are making an absolute claim. The account of hedging is also compatible with evidence that hedging can improve

[7] Intuitively, example (a) - "sort of stupid" - appears to work differently, by adjusting the meaning of "stupid", say, to "stupid in some relevant respect". We thank an anonymous reviewer for this point.

perceptions of a source's credibility (Jensen, 2008; though see Longman, Turner, King, & McCaffery, 2012), and that using some kinds of probabilistic language can protect a speaker's reputation (Teigen, 1988). To the best of our knowledge, however, no study has directly tested whether evidential hedging protects reputations.

Overview of the Experiments

In this paper we report six novel experiments that treat the following question: does hedging allow speakers to make low expectedness (low prior) claims without damaging their reputation? We try to answer this question for cases of pure expectation and cases when the outcome is known. Throughout we operationalize a speaker's reputation as their perceived reliability as measured on a ratings scale. The first four experiments are variations on the same fundamental design; we report these as a mini meta-analysis. These experiments explored hedging using a propositional attitude ("I suspect") or a double hedge ("I might be wrong, but I think..."). We then report two further experiments that extend the results of the meta-analysis.

As a whole, the experiments test between two contrasting sets of predictions, termed, here, the *Vigilance Position* and the *Hedging Position*. The *Vigilance Position* is inspired by expectation-based updating; the *Hedging Position* is inspired by McCready's (2015) account of hedging:

Vigilance Position: Reputation updating is robust to hedging.

Weak Form: Low expectedness claims cause perceived reliability to decrease even in the presence of hedges.

Strong Form: Hedging offers no protection; there are no differences between hedged and unhedged claims.

Hedging Position: Hedging protects against reputation updating

Weak Form: Hedged low expectedness claims give rise to better ratings of perceived reliability than unhedged low expectedness claims.

Strong Form: Hedging offers complete protection; hedged low expectedness claims cause no decrease in perceived reliability.

The weak forms of the *Vigilance* and *Hedging Positions* are compatible (in as much as vigilance may decrease perceived reliability, but hedging cushions that blow to speaker reputation); otherwise, the positions conflict. *Strong Vigilance* implies the negation of *Weak Hedging*: it is equivalent to a null hypothesis for hedging. *Strong Hedging* implies the negation of *Weak Vigilance*: it is equivalent to a null hypothesis for vigilance. In this paper, then, we focus on *Weak Vigilance* and *Weak Hedging*.

Later, in Experiments 5 and 6, we also test for a difference between hedging with propositional attitudes and double hedging. On a plausible reading of McCready (2015), hedging with propositional attitudes may offer some protection against a reputational hit by weakening the assertion. But double hedging should be more effective because of the disclaimer "I might be wrong, but..." Hence, double hedging should lead to higher ratings of perceived reliability than propositional attitudes.

To preview the results, across the experiments there was no clear evidence for a reliable effect of hedging. Perceived reliability was lower in the presence of low expectedness (low prior) claims, whether hedged or unhedged; and there were no reliable differences between hedged and unhedged claims or between hedging with propositional attitudes and double hedging.

Mini Meta-Analysis: Experiments 1 to 4

We begin our test of the *Vigilance* and *Hedging Positions* with a mini meta-analysis of four experiments[8]. These experiments followed the same basic design, and were simple belief-change tasks based on Collins, Hahn, von Gerber, and Olsson (2015, 2018), who set out to test predictions from two Bayesian models of testimony: the models of Bovens and Hartmann (2003) and Olsson and Angere (reported in Olsson, 2011; Olsson & Vallinder, 2013). Most relevant for present purposes is that, as the models predicted, low expectedness (low prior) information decreased the perceived reliability of its source (in Experiment 2 in Collins et al., 2015, which is Experiment 1b in Collins et al., 2018). In other words, low expectedness information damaged its source's reputation.

To introduce the present experiments, we consider, first, the basic method used by Collins et al. (2015, 2018): a pre-test/post-design design. Participants read an initial description of a source[9], whose reliability they then rated on a scale from 0 (not at all reliable) to 10 (completely reliable). For instance, participants saw the following item:

Michael is a clinical nurse specialist.

How reliable do you think he is, from 0 (not at all reliable) to 10 (completely reliable)?

[8] For brevity's sake we compress some details. We note, also, that Experiments 1 to 3 also included a condition in which claims were embedded under the phrase "I am certain that" so that we could explore the effect of propositional attitudes more generally. For fuller details and data on these additional conditions, see the unpublished PhD dissertation of Collins (2017).

[9] In other experiments, Collins et al. (2015, 2018) measured belief change instead of source reliability. Initially participants read a claim, rated their belief in it, and then saw the claim again, this time uttered by a source.

They then read the source making a claim, after which they rated the source's reliability again on the same scale. Some participants saw a high expectedness claim, that is, a claim with intuitively high prior probability:

> Now imagine that Michael told you the following: 'One of the best remedies against a severe cough is lots to drink, hot or cold.'

Other participants saw a low expectedness claim, that is, a claim with intuitively low prior probability:

> Now imagine that Michael told you the following: 'One of the best remedies against a severe cough is valium.'

In each case, participants answered the following question[10]:

> *Now how reliable do you think he is, from 0 (not at all reliable) to 10 (completely reliable)?*

Participants' prior ratings were subtracted from their posterior ratings to create change scores. Positive change scores meant an increase in perceived reliability; negative change scores, a decrease in perceived reliability. In an initial data set, high expectedness claims lead to a statistically reliable increase in perceived reliability (reliably positive change scores); low expectedness claims lead to a statistically reliable decrease in perceived reliability (reliably negative change scores). A replication yielded the same pattern for high expectedness claims, but the pattern was different for low expectedness claims. There was a trend in the anticipated direction, but this proved statistically inconclusive.

[10] We note that this question is somewhat vague: reliable about what topic and on which occasion? Experiments 1 to 4 use this vague question, but Experiments 5 and 6 insert a topic to address this issue.

With some adaptations - and some replication - the paradigm above offers a potential method for testing between the *Vigilance* and *Hedging Positions*. The same basic adaptations feature in Experiments 1 to 4. The experiments used only the low expectedness claims. Since, in Collins et al. (2015, 2018), participants gave these claims low initial ratings of convincingness[11] - in other words, low prior probabilities - these claims are an appropriate test case. The experiments modified these claims with (a) hedge(s), and used only reliable sources to avoid any floor effects. The experiments still followed a pre-test/post-test format, with participants rating sources' reliability before and after seeing the source make the low expectedness claim. The key manipulation was hedging: whether the claim was unhedged - a bold assertion - or hedged. This manipulation was between-participants.

Experiments 1 to 4 had a number of aims, including exploring the effects of propositional attitudes more generally. For present purposes we focus on hedging with the weak propositional attitude 'suspect' (see Methods for justification of that choice), as in Experiments 1 to 3, and the double hedge 'I might be wrong, but I think...'), as in Experiment 4.

Now that the design is clear, we can address some assumptions we made to generate predictions for Experiments 1 to 4, which were necessary because there is an imperfect fit between expectation-based updating and McCready's (2015) hedging. Firstly, McCready takes reliability to be induced from a history of interactions. In this

[11] The relevant data are from Experiments 1a and 2a in Collins et al. (2018). Pooling the data for both experiments (as is legitimate in a Bayesian analysis) for the claims we include in the present paper, a Bayesian one-sample *t*-test in the BEST package (Meredith & Kruschke, 2013) showed a mean estimate rating of 3.62, 95 % HDI[3.23, 4.01]. Hence the claims were rated comfortably below the mid-point of the scale (0 to 10).

and all other experiments, we instead used background knowledge or context to suggest sources' reliability, based on previous research (e.g. Collins et al., 2018). We assumed that this is a reasonable simplification. Secondly, McCready focuses on the case where "hedges shield the speaker from blame [reputational damage] if it turns out that her assertion fails to represent the facts correctly"(McCready, 2015, p. 3). In Experiments 1 to 4 (and, later, in Experiment 5) we, instead, focussed on expectations: participants never learnt the truth of the matter. We took expectations to set the bar lower. However different outcomes and expectations might be from a rational point of view (see the distinction about accuracy in the General Introduction), from the recipient's point of view there seems to be a continuum between things they know to be false and things they expect to be false. If a recipient finds a claim implausible, they presumably assign it a low probability; if a recipient learns that claim is false, they presumably assign it zero probability[12]. It is unclear why hedging should work in the more extreme, zero-probability case, and not the low-probability case. The less extreme case seems, to us, an easier target. We will return to this point, though, in Experiment 6 and the General Discussion.

These assumptions having been made, the predictions were as follows. This was a single-factor between-participants design, with the variable Hedging with levels 'Unhedged' and 'Hedged-Suspect' (Experiments 1 to 3) or 'Hedged-Double' (i.e. 'I might be wrong, but I think...'; Experiment 4). The pre-test/post-test structure allowed various comparisons: one-sample comparisons (versus zero) for each condition, and comparisons between hedged and unhedged conditions. We predicted the following.

[12] Or near zero: a recipient might want to acknowledge some small chance of error.

Basic Reputational Hit: in the unhedged condition, perceived reliability should reliably decrease (one-sample comparison).

Weak Vigilance: perceived reliability should reliably decrease in the unhedged, hedged-suspect and hedged-double conditions (one-sample comparison).

Weak Hedging: in Experiments 1 to 4 change scores should differ reliably between the unhedged and hedged-suspect and the unhedged and hedged-double conditions, with the unhedged condition reliably lower than both of the hedged conditions, indicating a bigger decrease in perceived reliability (between-groups comparison)

Methods

Participants. Table 1 summarizes the number of participants for each experiment. The numbers reported are for the experiment as a whole, including the conditions not reported in this paper. Sample sizes were selected to give high power (.8) for a medium effect size in frequentist analyses (the original analyses were frequentist). Numbers are reported for the original samples and for the samples retained after exclusions. For all experiments, participants were excluded if they stated that their first language was not English; for Experiments 2 to 4, they were also excluded if they failed a memory check or indicated that they did not participate seriously (see Procedure). Gender and mean age are reported for the participants whose data were retained.

Table 1. Participants for Experiments 1 to 4

Experiment	N(Sample)	N(Retained)	N(Female)	Mean Age
1	160	159	84	37.35
2	170	165	96	39.93
3	161	151	75	37.69
4	128	124	71 (and 1 non-binary)	38.36

Materials. There were five items, which took the following form:

Initial source information: Michael is a clinical nurse specialist.

Claim: Now imagine that Michael told you the following: 'One of the best remedies for a severe cough is valium.'

The item above corresponds to the unhedged condition. In the hedged condition, participants read the claim embedded under "I suspect that..."

We adopted the items from Collins et al. (2018), selecting reliable sources making low expectedness claims. A clinical nurse specialist claimed that valium was one of the best remedies for a severe cough. A baker claimed that the varying temperatures of a particular oven made it perfect for fluffy and crispy bread. A journalist with a good track record predicted that a horse would beat a competitor despite losing to that horse in the majority of races. A retired meteorologist claimed that the maximum summer temperature in Stockholm in June 2013 was 15 degrees.

And a respected DJ claimed that a club in Detroit had the reputation of one of coolest in the world.

The hedged conditions embedded the claims underneath the propositional attitude 'suspect' (hedged-suspect; Experiments 1 to 3) or the double hedge 'I might be wrong, but I think...' (hedged-double; Experiment 4). Both types of hedge feature in McCready's analysis. In the present study, Experiments 3 and 4 also included manipulation checks in which participants clearly found hedged claims weaker than their unhedged counterparts: in Experiment 3 "suspect" was rated on average 4.92 (*SD* = 2.04) on a scale from 0 (very weak) to 10 (very strong), whereas the unhedged claims were rated on average 7.70 (*SD*= 2.43); in Experiment 4, "suspect" was rated on average 5.27 (*SD* = 1.91) and "I might be wrong, but..." was rated on average 4.96 (*SD* = 1.93), whereas the unhedged claims were rated on average 8.02 (*SD* = 2.39). The hedged (hedged-suspect and hedged-double) claims were rated as conveying significantly weaker claims than unhedged claims (*ps* < .001; for details, see Collins, 2017).

Procedure. The experiments reported here and below were carried out following the guidelines, and with the approval, of the ethics committee at the Department of Psychological Sciences, Birkbeck, University of London. The experiments were posted on Amazon Mechanical Turk via the intermediary MTurk Data (www.mturkdata.com). We set high qualifications for the task: participants had to be resident in the US and have an overall approval rating of 99% for 1,000 previously completed tasks. These criteria maximized the number of first-language English speakers. Each participant was only able to take part in a single experiment. On the consent page, participants were told that the task would assess how people judge

information they receive from other people. No information was given about the specific manipulation. After giving informed consent, participants were assigned, in a round-robin manner, to a condition. They were shown the following instructions:

> Thank you for taking part in this study. You'll be shown some descriptions of people, and will be asked to indicate how reliable these people seem, by selecting the appropriate number on a scale from 0 (not at all reliable) to 10 (completely reliable).

Participants then saw 5 items, comprising initial source information and then a claim. Participants gave a prior rating after the initial source information, and a posterior rating after the claim. The order of presentation was counterbalanced. Finally, participants were given debriefing information, and were paid $0.75, a fee chosen to exceed the US minimum wage for reasonable completion times.

In Experiments 2 to 4, there were the following changes to the procedure. To ensure that participants processed the materials deeply enough, participants were informed that they may have to perform a memory test. Participants in the hedged conditions did, indeed, perform a memory test at the end of the experiment and in a fixed order. The test comprised a recall test and a recognition test, in that order, details of which can be found at osf.io/r9cna. The recall test gave an indication of how deeply participants were processing the materials: for instance, in Experiment 2, approximately 76% passed the test, rising to 84% if near synonyms are included. But the recognition test was the minimum threshold for inclusion of the data. This threshold was waived if participants passed the recall test but failed the recognition test, since these participants may have misinterpreted the recognition question as a cue that their recall was incorrect. Then, participants answered a sincerity question,

based on Aust, Diedenhofen, Ullrich, and Musch (2013); for details see osf.io/r9cna. Experiments 2 and 3 also placed the propositional attitude 'suspect' in capitals as a simple attentional device; this device was dropped in Experiment 4. And Experiments 3 and 4 included a manipulation check between the recognition test and the sincerity question; for the wordings, see osf.io/r9cna.

Results

Copies of the data files and analysis scripts are available on the Open Science Framework at osf.io/r9cna. For all data, we calculated change scores by subtracting the prior rating from the posterior rating. We then averaged change scores across items to produce a single score per participant[13]. For the present meta-analysis, we analysed the data in two ways: we first calculated meta-analytic means and confidence intervals in R (R Core Team, 2016) using the 'meta' package (Schwarzer, 2007). Since we are only interested in summarizing the present studies, we used a fixed-effects model. We then entered the relevant *t*-statistics from the experiments into a meta-analytic Bayesian *t*-test in the 'Bayes Factor' package (Morey & Rouder, 2015). Bayes Factors allow us to directly assess the support for the null and alternative hypotheses. We assess each of the predictions in turn.

Basic Reputational Hit. Fig. 1 is a forest plot of the mean change scores and confidence intervals for the unhedged condition across the four experiments (studies). The figure shows reliably negative change scores: participants reliably decreased their

[13] Given this averaging, we must be cautious about generalizing beyond the existing items, but see Experiment 5.

belief in the reliability of the source when s/he made a low expectedness claim. Hence, the data replicate the basic reputational hit.

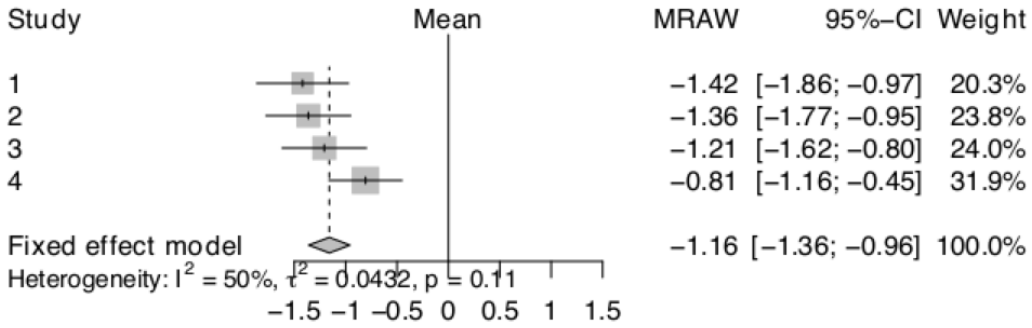

Figure 1. Forest plot of mean change scores and confidence intervals for the unhedged condition of Experiiments 1 to 4.

A meta-analytic Bayesian t-test confirmed this result. Table 2 reports the one-sample *t*-statistics, number of participants, and degrees of freedom for the unhedged condition in each study.

Table 2. One-sample *t*-statistics, degrees of freedom, and number of participants for the unhedged condition for Experiments 1 to 4.

Study	*t*-statistic (df)	*N*
1	-6.23 (53)	54
2	-6.53 (58)	59
3	-5.83 (51)	52
4	-4.43 (57)	58

The Bayesian analysis found strong support for a reputational hit from low expectedness claims. To gauge sensitivity to priors, the analysis used the three different pre-set Cauchy priors in the Bayes Factor package for the standardized effect

size *r*: medium, wide, and ultra-wide. These each yielded strong support for a reputational hit, $BFs_{Alternative} > 100$.

Weak Vigilance. Fig. 2 is a forest plot of the mean change scores and confidence intervals for the hedged condition across the four experiments (hedged-suspect in Experiment (Study) 1 to 3; hedged-double in Experiment (Study) 4). The figure shows reliably negative change scores: participants decreased their belief in the reliability of the source when s/he made a low expectedness claim even when that claim was hedged.

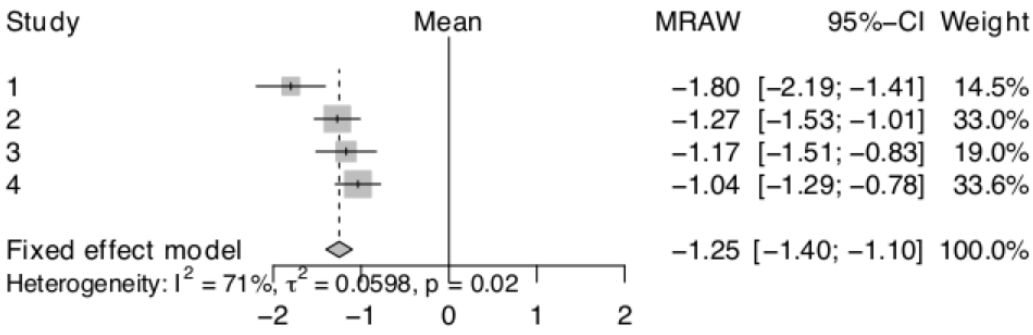

Figure 2. Forest plot of mean change scores and confidence intervals for the hedged condition for Experiments 1 to 4.

A meta-analytic Bayesian *t*-test confirmed this result. Table 3 reports the *t*-statistics, number of participants, and degrees of freedom for the hedged condition.

Table 3. One-sample *t*-statistics, degrees of freedom, and number of participants for the hedged condition for Experiments 1 to 4.

Study	*t*-statistic (df)	*N*
1	-8.99 (53)	54

2	-9.66 (51)	52
3	-6.64 (46)	47
4	-7.89 (66)	67

The Bayesian analysis found strong support for a reputational hit with hedged claims, hence for *Weak Vigilance*. As above, the analysis used the three pre-set Cauchy priors, which each yielded strong support for *Weak Vigilance*, $BFs_{Alternative} > 100$.

Weak Hedging. Fig. 3 is a forest plot summarizing the difference in change scores between the unhedged and hedged conditions across the four experiments (studies). For each study, the confidence intervals include 0; the meta-analytic confidence interval also includes 0, and the meta-analytic mean is slightly negative, with non-significantly lower ratings in the hedged conditions. The figure offers no support for the predicted effect of hedging.

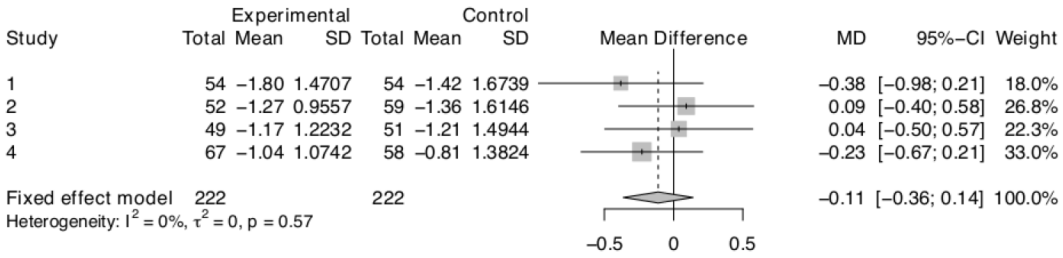

Figure 3. Forest plot of differences in mean change score between hedged and unhedged conditions (and confidence intervals) for Experiments 1 to 4.

A meta-analytic Bayesian t-test confirmed this result. Table 4 reports the independent-samples *t*-statistics, number of participants in hedged and unhedged groups, and the degrees of freedom across the four experiments.

Table 4. Independent-samples *t*-statistics, degrees of freedom, and participant numbers for difference in mean change scores for hedged and unhedged conditions for Experiments 1 to 4.

Study	*t*-statistic (df)	*N*(Hedged)	*N*(Unhedged)
1	1.26 (104.27)	54	54
2	-.33 (96.41)	52	59
3	-.14 (95.88)	47	52
4	1.02 (106.73)	67	58

The Bayesian analysis found moderate support for the null hypothesis, hence moderate evidence against *Weak Hedging*. As above, the Bayesian analysis used the three pre-set Cauchy priors for the effect size. In each case there was moderate support for the null hypothesis, but there was some sensitivity to those priors: with the default medium prior, $BF_{Null} = 6.23$; with the wide prior, $BF_{Null} = 8.68$; and with the ultra-wide prior, $BF_{Null} = 12.18$.

Discussion

The mini meta-analysis is informative on a number of points. Firstly, it replicated the basic reputational hit from low expectedness claims which was found in the initial study by Collins et al. (2015, 2018) but which was inconclusive in a replication (Collins et al., 2018). This finding is consistent with Bayesian models of testimony (Bovens & Hartmann, 2003; Olsson, 2011; Olsson & Vallinder, 2013) and with the idea of epistemic vigilance (Sperber et al., 2010). The finding generalizes, moreover, to low expectedness claims which are embedded under hedges. The data

support at least *Weak Vigilance*: low expectedness claims cause a reputational hit even when hedged. The data do not support *Weak Hedging*: hedging did not reliably protect reputations. Indeed, the data suggest moderate evidence for the relevant null hypothesis. In other words, epistemic vigilance seems robust against hedging. The data do not sit well with the idea that a key function of hedges is the protection of reputation, in the sense of perceived reliability (McCready, 2015).

The results above arise from experiments with slightly different procedures. Experiments 2 to 4 sought to ensure that the hedges were not simply ignored. Experiments 2 and 3 placed the hedge, 'suspect', in capitals; Experiments 2 to 4 all included a memory test, which encouraged participants to process the text more deeply; and Experiments 3 and 4 also included manipulation checks. These procedural differences make no obvious difference to the results, and suggest that participants in all experiments processed the items in some depth.

Nevertheless, Experiments 1 to 4 have a number of limitations. The experiments used a design and statistical analysis that limit generalizability. While we might expect our results to generalize to new participants, we have no grounds to expect them to generalize to new materials (Clark, 1973). The experiments also used a pre-test/post-test format, a method that may give rise to demand characteristics (Levine & Parkinson, 1994). In particular, participants may have deduced that we intended them to use the content of the claims to revise their judgments of source reliability. Since the hedging manipulation was between-participants, participants had no evidence to guess the manipulation itself. But the results are worth replicating with an alternative, more natural design. Finally, the experiments did not include any definition of 'reliability' for participants; we assumed, instead, that participants would

use an intuitive notion. This intuitive notion may not, however, map neatly onto the notion of reliability that is theoretically relevant.

Experiment 5

Experiment 5 set out to address the limitations of the preceding studies. It used a design that yields better evidence of generalizability, addresses potential demand characteristics of the pre-test/post-test format, and tightens up the dependent measure.

The basic design was as follows. The experiment included 20 items in total, including adapted versions of the materials from Experiments 1 to 4. To reduce task demands and vary demand characteristics, Experiment 5 abandoned the pre-test/post-test format, and adopted a fully between-participants design which comprised the following 4 conditions.

Null Condition: participants read only information about 20 sources.

Unhedged Condition: participants read these 20 sources making low expectedness claims (one claim per source).

Propositional-Attitude Condition: participants read these 20 sources making the same low expectedness claims but hedged with a propositional attitude ('I suspect that...').

Double-Hedged Condition: participants read these 20 sources making the same low expectedness claims but hedged with 'I might be wrong, but I think...'

Participants rated source reliability, this time using the (potentially) more sensitive measure of fixing a slider on a scale from 0 to 100 (on the definition, see below).

In this design, the null condition provides a baseline for the reliability of a source. It anchors the reputational hit: a reputational hit occurs in any condition where ratings are reliably lower than the null condition. The basic effect of low expectedness claims should be seen in the difference between null and unhedged conditions. The effects of hedging can be seen in the difference between unhedged and hedged (Propositional-Attitude, Double-Hedged) conditions. If hedging protects the source's reliability, then ratings should be reliably higher in hedged conditions than in the unhedged condition. This design also permits direct comparison between hedging with a propositional attitude and double hedging.

The new design has certain advantages over the preceding studies but prompts further changes. First, consider some advantages. Since the design is fully between-participants, it reduces (or at least varies) the potential demand characteristics. In the between-participants design, participants give a single rating after reading the items, and only a control group (the Null Condition) rates the source when s/he has not made a claim. The design also includes more items, allowing us to be more confident about generalizing our findings.

The fully between-participants design might, however, alter how people understand the task. Accordingly we introduced new features to protect construct validity. These features largely concern the dependent variable, perceived reliability. Since different conditions present different amounts of information, participants in these conditions might interpret reliability differently. In particular, participants in the null condition do not see a claim at all, which risks them giving more general ratings than participants in the other conditions. It is one thing to rate a nurse's reliability as a source in general and another to rate the nurse's reliability as a source about medical

matters. To aid interpretation, we made the following four changes across all conditions. Firstly, we defined reliability as a person's "credibility, trustworthiness, or expertise - as whether they tend to say the truth". Including this definition also mitigates the concern that participants might understand reliability differently from the theoretically important sense. Secondly, we added information to each item to provide a context for the rating. So, for example, the new version of the now-familiar Michael item was as follows, with the relevant information italicized:

> *You are discussing treatments for coughs and colds,* which are mostly treated with simple over-the-counter medicines. Michael is a nurse practitioner who specializes in minor illnesses.

Thirdly, we fixed the scope of the reliability question for each item by identifying a topic. For the Michael item (topic italicized):

> How reliable is Michael *about remedies for minor illnesses*?

Since there is no objective measure of topic scale - Is the evolution of the domestic cat a bigger/smaller topic than the history of the Napoleonic wars? - we relied on intuition to keep the scale comparable. In each case, the topic was broad enough to avoid collapsing to a simple judgment of belief in a claim in the conditions where a claim was presented (unhedged, propositional attitudes, and double hedged). Fourthly, we gave participants three practice items to sensitize them to the full range of the scale. These items described a high reliability (highly trustworthy and expert) source, a middle-reliability source (with a 50:50 hit rate), and a low reliability (highly inexpert) source.

A final innovation concerns expectedness. In Experiments 1 to 4, the items varied in how expectedness was determined. In some items, expectedness was

determined by (presumed) background beliefs: for instance, we presumed that participants would have a sense of likely and unlikely summer temperatures. In other items, expectedness was determined by contextual information: for instance, we told participants about the record of a pair of horses in races against each other. In Experiment 5, when items invoked background knowledge, they also explicitly stated that knowledge: hence, for example, an item about summer temperatures in North Africa explicitly stated that summers there are very hot.

For this design we make the following predictions:

Reputational Hit: the Unhedged Condition should have reliability ratings reliably lower than the Null condition.

Weak Vigilance: reliability ratings should be reliably lower than the Null Condition in all other conditions.

Weak Hedging: the hedged conditions (Propositional Attitude and Double-Hedged Conditions) should have reliability ratings reliably higher than the Unhedged Condition.

Shielding: the Double-Hedged condition should be rated reliablly higher than the Propositional Attitude condition.

Methods

Participants. 206 participants completed a web survey; the same selection criteria were used as for the previous experiments. We retained the data for the 203 participants (88 female; average age 35.98) who indicated that English was their first language.

Materials. There were 3 practice items (see Procedure) and 20 test items. These can be seen in a copy of the Qualtrics survey on the Open Science Framework at osf.io/r9cna. Each test item described a scenario, gave information about a source, and set up an expectation. The Null condition comprised only this information. In the Unhedged, Propositional-Attitude and Double-Hedged conditions the source also made a low expectedness claim (given the context). In the hedged conditions, the claim was modified by a propositional attitude ('I suspect that...') or a double hedge ('I might be wrong, but I think that...'). Finally, there was a reliability question: 'How reliable is [Source] about [Topic]?' The following is one of the items. In the Null condition, participants saw only the context and the reliability question.

> You are talking about Susie, a high-school student, and her work habits. Susie is known for being easily distracted. Jake is Susie's brother, and they both still live at home.
>
> *Unhedged*: He tells you, 'Susie does her homework best sitting in front of the TV.
>
> *Propositional Attitude:* He tells you 'I suspect that Susie does her homework best sitting in front of the TV.
>
> *Double-Hedged*: He tells you, 'I might be wrong, but I think that Susie does her homework based sitting in front of the TV.'
>
> How reliable is Jake about Susie?

Procedure. The experiment comprised a Qualtrics web-survey, posted on Mechanical Turk by the intermediary MTurkData. Participants gave informed consent and were (pseudo-)randomly assigned to a condition. All participants received the following instructions:

In this task, we would like you to rate how reliable you think certain people are as sources of information. You can think of someone's reliability as their credibility, trustworthiness, or expertise - as whether they tend to say the truth.

You'll see 20 scenarios. For each one you'll be asked about someone's reliability. We would like you to make your judgments using the sliding scale provided. '0' means not at all reliable, '100' completely reliable. You now have the chance to practice on the following items.

They then rated the following items using a sliding scale from 0 to 100, to sensitize themselves to the scale.

[High Reliability] You are being told about a particular type of deep-sea fish by a world expert on the subject. The expert has no reason to misinform you. How reliable is the expert about this fish?

[Medium Reliability] You are being told by a high-school student about what they learnt in class today. The student has a vivid imagination, and makes things up about half the time. How reliable is the student about their class?

[Low Reliability] You are visiting a foreign city. Another tourist is telling you about the city. The tourist has only briefly looked at a guidebook in a foreign language that they understand very poorly. How reliable is the tourist about the city?

Participants were then reminded of the instructions using the following text:

Thank you. Let's get on with the rest of the experiment. Now you'll see 20 items, each about a different topic. Just as with the practice questions, you'll

be asked about the reliability of some people as sources of information about a topic. As you've just practiced, please rate their reliability using the scale provided. '0' means not at all reliable. '100' means completely reliable.

Each participant then rated 20 items. The web survey randomized the order of presentation. Finally, participants received debriefing information, and received a fee of $1.50, which was chosen to exceed the US minimum wage per minute.

Results

Copies of the data files and analysis scripts can be found on the Open Science Framework at osf.io/r9cna.The analysis comprised a Bayesian zero-one inflated beta regression run in the 'brms' package (Bürkner, 2017). Conclusions, here, are based on the estimated marginal means and the 95% Highest Posterior Density intervals, calculated with the emmeans package (Lenth, 2018). A survey error led to ambiguities in two items (Items 11 ("Eric") and 12 ("Alex") in the survey; see the Open Science project) that may have decreased the effect of hedging. Accordingly we excluded these items, and ran all analyses on the remaining 18 items.

Fig. 4 shows the descriptive data and was produced using the 'pirateplot' function of the 'yarr package' (Phillips, 2017). The descriptive data suggest a clear drop in perceived source reliability between the Null and remaining conditions. There are small differences between the remaining conditions with the Propositional

Attitude (PropAtt) condition rated slightly worse than the Unhedged and Double-Hedged (Double) condition.

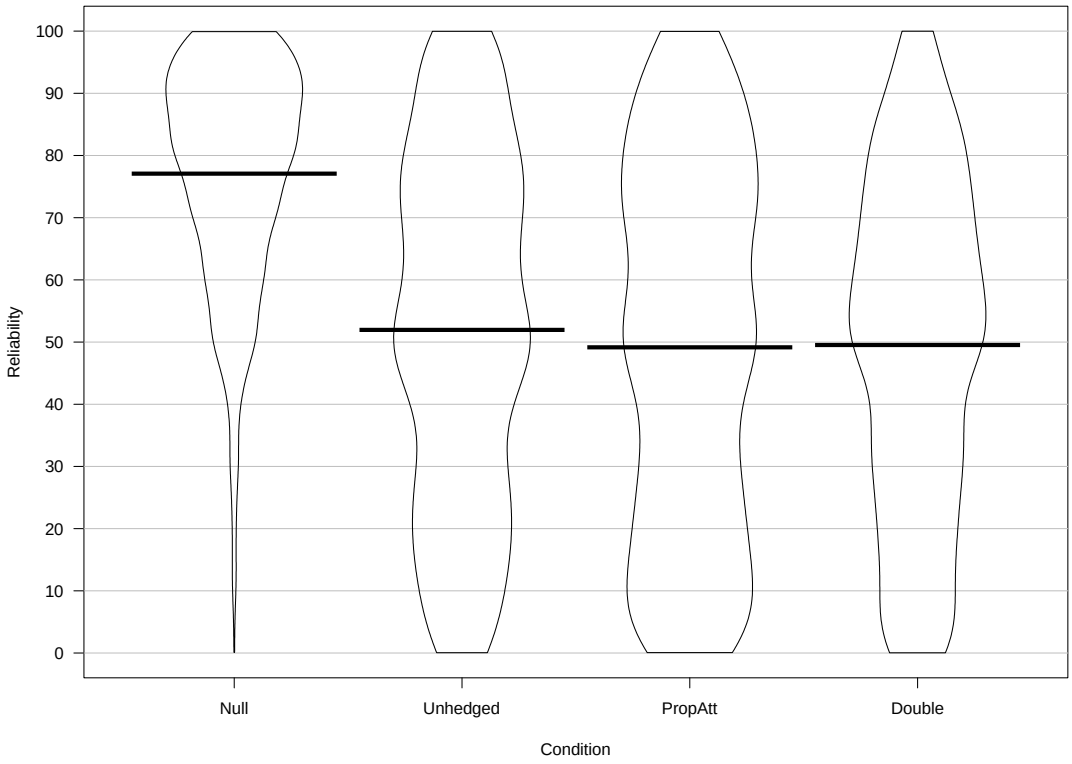

Figure 4. The figure shows the horizontally jittered raw data and smoothed density distributions by condition. Black bars represent the mean for each condition; the white boxes represent the 95% high density interval (see Kruschke, 2013).

We analyzed the data with a zero-one-inflated beta regression with perceived reliability (rescaled to the range 0-1) predicted by Condition (Null, Unhedged, Propositional Attitude, Double-Hedging). The model included random slopes and intercepts for items and random intercepts for participants: that is, in R syntax, the model included the terms '(Condition|Item)' and '(1|ID)', where 'ID' stands for 'participant'.

We used the default uninformative, or weakly informative, priors from the brms package. For the beta coefficients, the prior for the intercept was a 0-centered

Student-*t* distribution with 3 degrees of freedom and scale of 31; the prior for the remaining coefficients was an improper uniform prior across the real numbers. For the standard deviations of the random effects, the prior was a 0-centered Student-*t* distribution with 3 degrees of freedom and scale of 31); for the correlation between random effects, the prior, ('lkj(1)'), set all correlation matrices to be equally likely. For the conditional-one inflation (coi) and zero-one inflation probabilities (zoi), the priors were uniform (Beta($\alpha = 1$, $\beta = 1$). For phi, the prior was a gamma distribution with shape and spread of .01. We ran 3 chains for 10,000 iterations. These settings achieved good convergence for all chains. All Rhat values were 1.00; and effective sample sizes were above 3,800.

Table 5 reports the parameter estimates.

Table 5. Parameter estimates (excluding group-level parameters) and 95% Credible Intervals. Note that estimates are on the logit scale, and that the independent variable was treatment-coded.

Parameter	Parameter Estimate	95% Credible Interval
Null (Intercept)	1.05	.82, 1.27
Unhedged	-1.03	-1.31, -.75
Propositional Attitude	-1.10	-1.38, -.83
Double	-1.09	-1.36, -.83
phi	4.40	4.20, 4.61
zoi	.06	0.05, 0.07
coi	.65	.58, .71

Table 6 reports the estimated marginal means, which are on the response scale.

Table 6. Estimated marginal means and 95% HPDIs. Note that all estimates are on the original response scale.

Parameter	Estimated marginal mean	95% HPDI
Null (Intercept)	74.13	69.62, 78.25
Unhedged	50.55	43.38, 57.89
Propositional Attitude	48.87	42.07, 55.78
Double	49.05	42.75, 55.60

The estimated marginal means show that the experimental (non-null) conditions are clearly reliably lower than the Null Condition, but are themselves close together with substantial overlap in the 95% highest posterior density intervals. We consider the evidence for each hypothesis in turn.

Reputational Hit. The data replicate the basic reputational hit seen in Experiments 1 to 4: ratings are reliably lower in the Unhedged Condition than in the Null Condition.

Weak Vigilance. The data also replicate the finding that ratings are reliably lower in both the Propositional Attitude condition and the Double Hedged condition than in the Null Condition.

Weak Hedging. The data show no clear evidence for an effect of hedging: the hedged conditions do not have ratings reliably higher than the Unhedged Condition: there is almost complete overlap in the HPD intervals, and the mean ratings were slightly lower for hedged conditions than for the Unhedged Condition.

Shielding. The data show no evidence of shielding: the Double-Hedged condition did not have reliability ratings reliably higher than the Propositional Attitude Condition. Indeed, there was almost complete overlap of 95% HPD intervals, for the two types of hedging: the Double-Hedged Condition extended only .44 points on the reliability scale above the Propositional Attitude Condition.

We tested model fit with posterior predictive checks, shown in Fig. 5. These checks suggest adequate fit, though note the mismatch between approximately .25 and .50 (25 and 50 on the response scale): the model predicts too many ratings in this interval.

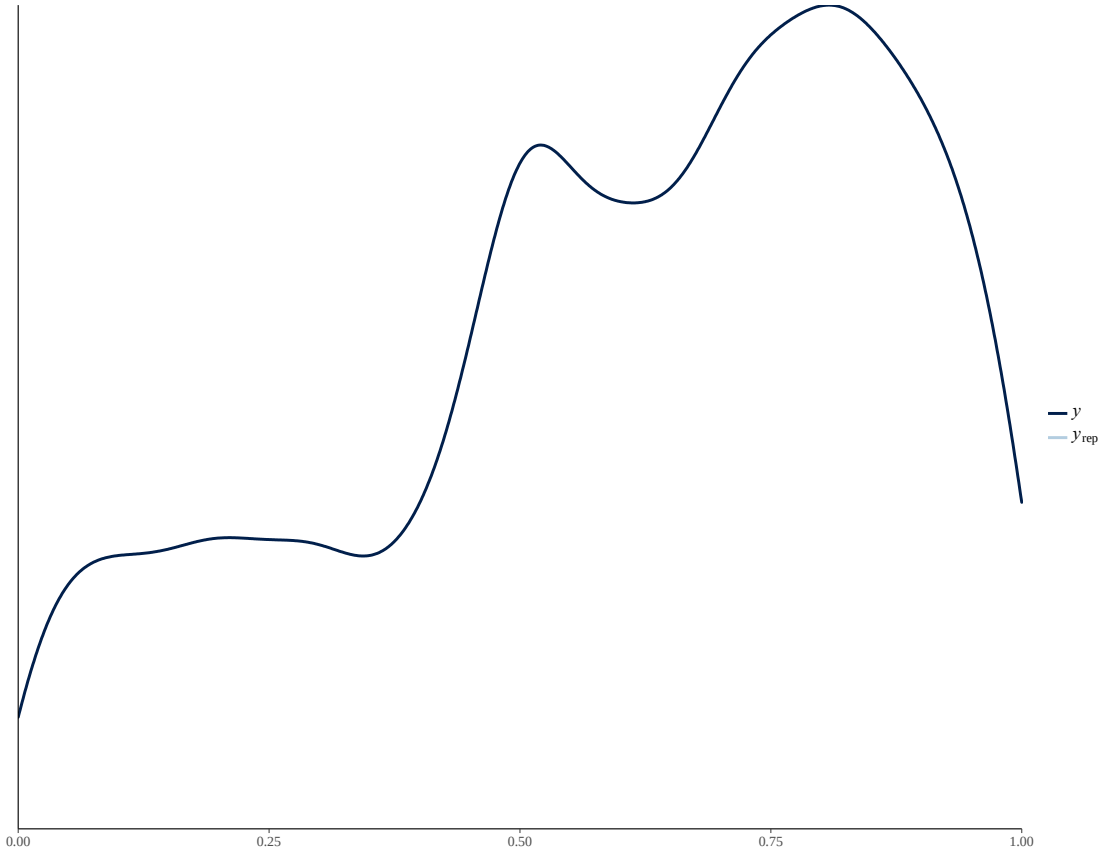

Figure 5. Posterior predictive checks for the model, showing the data (y) against 100 predictions from the model (yrep).

In summary, the data for Experiment 5 only offer support for *Weak Vigilance*. They suggest, further, that ratings were highly similar for the Unhedged, Propositional Attitude, and Double Hedged Conditions.

Discussion

A fifth experiment has now replicated the reputational hit from low expectedness claims and shown that it can occur even when the claims are hedged. This experiment used new items, a new design, and a more tightly defined dependent variable. Nevertheless, the data replicate the basic pattern seen in Experiments 1 to 4: support for *Weak Vigilance* and no support for *Weak Hedging*. Experiment 5 adds a direct comparison between hedging with the propositional attitude 'I suspect that' and double-hedging with 'I might be wrong, but I think...' There was no reliable difference between these types of hedging.

At this point, however, we must reconsider our auxiliary assumptions and, in particular, the assumption that expectations are a fair test case for reputation management. The assumption seems reasonable: low expectedness claims seem to provide a clear case of reputation damage and should, on McCready's (2015) account, threaten cooperation. As we have already discussed, expectations seem to set the bar somewhat lower. And many conversations will be about things whose truth we cannot - at least not immediately - establish. Our approach, then, seems to treat an important part of the picture.

Nevertheless, McCready does not treat expectations, but cases where the outcome is known: to repeat an earlier quotation, she focuses on the case where "hedges shield the speaker from blame [reputational damage] if it turns out that her

assertion fails to represent the facts correctly"(McCready, 2015, p. 3). It might be a fairer test of McCready's account to include outcomes. Outcomes feature in our final experiment, to which we now turn.

Experiment 6

This experiment used the same design as Experiment 5 with only a single alteration. We added, to each item, the phrase 'Later you learn that what [source] told you was wrong'. We repeat the predictions from above:

Reputational Hit: the Unhedged Condition should have reliability ratings reliably lower than the Null condition.

Weak Vigilance: reliability ratings should be reliably lower than the Null Condition in all other conditions.

Weak Hedging: the hedged conditions (Propositional Attitude and Double-Hedged Conditions) should have reliability ratings reliably higher than the Unhedged Condition.

Shielding: the Double-Hedged condition should be rated reliably higher than the Propositional Attitude condition.

Methods

Participants. 205 participants completed a web survey; the same selection criteria were used as in all previous experiments. We retained the data for the 204 participants (86 female; average age = 36.86) who listed English as their first language.

Materials. The materials were the same as for Experiment 5 (the errors in items 11 and 12 having been corrected). For each item, an extra line was added: 'Later

you learn that what [source] told you was wrong.' A copy of the Qualtrics survey is available on the Open Science Framework at osf.io/r9cna.

Procedure. The procedure was identical to Experiment 5.

Results

Copies of the data files and analysis scripts are available on the Open Science Framework at osf.io/r9cna. As with Experiment 5, the analysis comprised Bayesian zero-one inflated beta regression run in the 'brms' package, and conclusions are based on the estimated marginal means and 95% HPDs.

Fig. 6 shows the descriptive data.

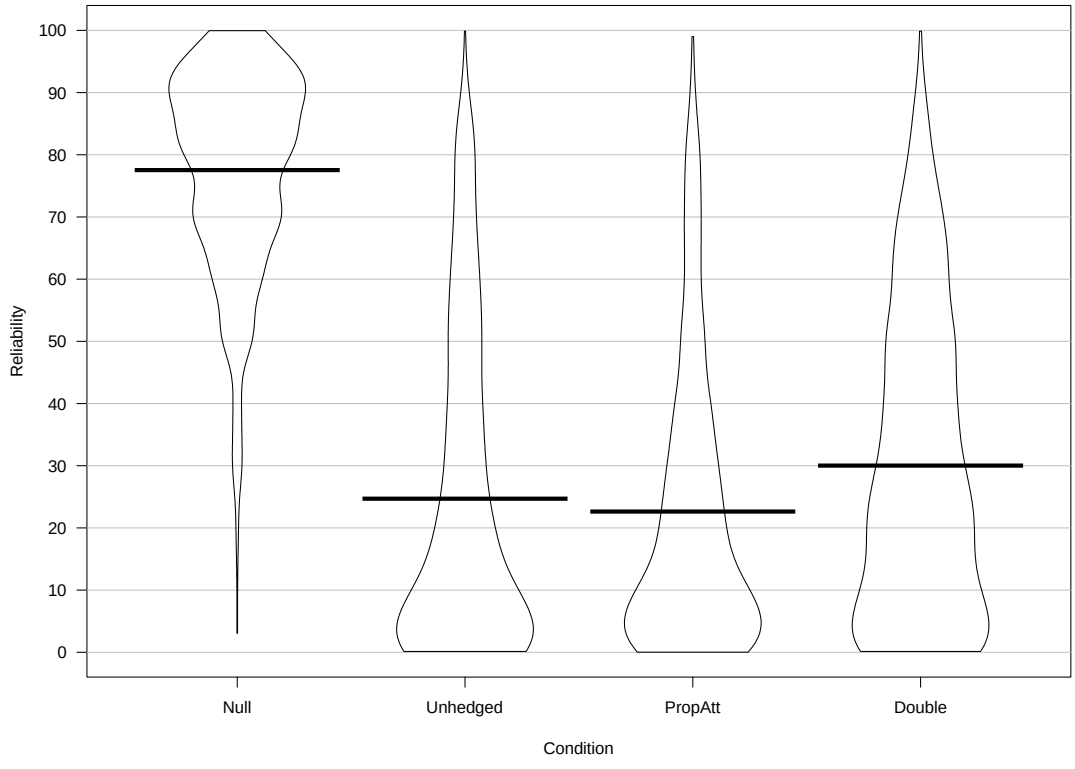

Figure 6. The figure shows the horizontally jittered raw data and smoothed density distributions by condition. Black bars represent the mean for each condition; the white boxes represent the 95% high density interval (see Kruschke, 2013).

The data suggest that, when low expectedness information proves false, there is a large hit to reliability. The pattern differs from that seen in Experiment 5: the Double-Hedging Condition has a noticeably higher mean than the Unhedged and Propositional Attitude Conditions.

We analyzed the data with a zero-one-inflated beta regression with perceived reliability (rescaled to the range 0-1) predicted by Condition (Null, Unhedged, Propositional Attitude, Double-Hedging). The model included random slopes and intercepts for items and random intercepts for participants: that is, in R syntax, the model included the terms '(Condition|Item)' and '(1|ID)', where 'ID' stands for 'participant'. We used the same priors as for Experiment 5, and once again ran 3 chains for 10,000 iterations, achieving good convergence. All Rhat values were 1.00; effective sample sizes were all above 1,100.

Table 7 reports the parameter estimates.

Table 7. Parameter estimates (excluding group-level parameters). Note that estimates are on the logit scale, and that the independent variable was treatment-coded.

Parameter	Parameter Estimate	95% Credible Interval
Null (Intercept)	1.22	.95, 1.48
Unhedged	-2.31	-2.65, -1.97
Propositional Attitude	-2.49	-2.82, -2.15
Double	-2.06	-2.38, -1.74
phi	7.78	7.42, 8.14
zoi	.11	.10, .12

coi	.10	.07, .13

Table 8 reports the estimated marginal means, which are on the response scale.

Table 8. Estimated marginal means and 95% HPDIs. Note that all estimates are on the original response scale.

Parameter	Estimated marginal mean	95% HPD
Null (Intercept)	77.15	71.20, 81.24
Unhedged	25.13	20.73, 30.53
Propositional Attitude	21.82	17.92, 26.56
Double	29.97	25.01, 35.70

The estimated marginal means show that the experimental (non-null) conditions are clearly reliably lower than the Null Condition. A new finding, here, is that the Double Hedging Condition has higher ratings than either the Unhedged or Propositional Attitude Conditions. However, there is overlap in the 95% HPD intervals, if rather minimal between the hedged conditions. Unsurprisingly, ratings in the non-null conditions are considerably, and reliably, lower in this experiment than in Experiment 5.

Reputational Hit. The data replicate the basic reputational hit seen in Experiments 1 to 4: ratings are reliably lower in the Unhedged Condition than in the Null Condition.

Weak Vigilance. The data also replicate the finding that ratings are reliably lower in both the Propositional Attitude condition and the Double Hedged condition.

Weak Hedging. The data do not show clear evidence of hedging. The Propositional-Attitude Condition once again has ratings lower than the Unhedged Condition, though these ratings are, once again, not reliably different. The Double Hedging Condition, this time, has higher ratings than the Unhedged Condition, but the 95% HPD intervals overlap by some 4.28 reliability points.

Shielding. The data hint at an effect of shielding. The Double-Hedged Condition had higher ratings than the Propositional Attitude Condition, but the 95% HPD intervals overlap by some 1.55 reliability points. There is, therefore, insufficient evidence to conclude in favour of an effect of shielding.

We tested model fit with posterior predictive checks, shown in Fig. 7. The checks suggest reasonable fit, though some over-prediction in the .75 to 1 range (75%

to 100% on the response scale) and some under prediction in approximately the .15 to .40 (15 to 40) range.

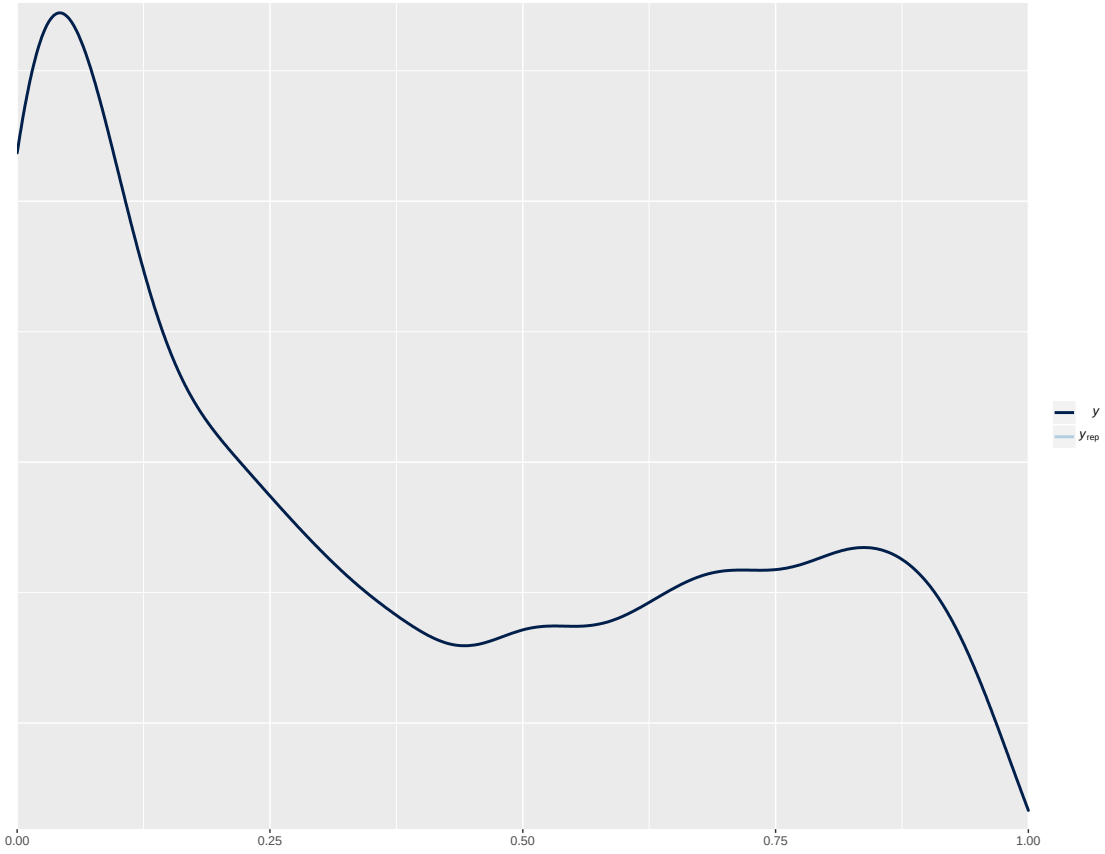

Figure 7. Posterior predictive checks, showing the data (y) against 100 simulations from the model (yrep).

In summary, the headline finding is that there is, once again, no clear evidence that hedging protects reliability. This headline should not, however, obscure the differences between the data sets. While Experiment 5 offered good evidence for the similarity of ratings in the Unhedged, Propositional Attitude, and Double Hedged Conditions, Experiment 6 offers tentative evidence for differences: in particular, between the Unhedged and Double Hedged Conditions - where the overlap in 95%

HPDIs is now much less - and between the Propositional Attitude and Double Hedged Conditions, where there is only small overlap.

We explored whether the above differences were statistically reliable by compiling the data sets for Experiments 5 and 6 and running a Bayesian mixed-effects zero-one-inflated beta regression. For this analysis, we excluded items 11 and 12 from Experiment 6 to match the exclusions from Experiment 5. The model included terms for Condition (Null, Unhedged, Propositional Attitude, and Double) and Experiment (Expectation, Outcome; where 'Expectation' stands for Experiment 5 and 'Outcome' for Experiment 6). We allowed the interaction and main effects to vary across items (i.e., in R syntax (Condition*Experiment|Item) and included random intercepts for items and participants. As above, we used the default priors in the brms package. We ran 3 chains for 10,000 iterations. These settings resulted in good convergence. All Rhat values were 1.00, and effective sample sizes were above 3,100.

Table 9 reports the parameter estimates.

Table 9. Parameter estimates (excluding the group-level parameters). Note that estimates are on the logit scale, and that both independent variables were treatment-coded.

Parameter	Parameter Estimate	95% Credible Interval

Null (Intercept)	1.08	.89, 1.27
Unhedged	-.98	-1.22, -.75
Propositional Attitude	-1.17	-1.40, -.95
Double	-1.12	-1.36, -.89
Experiment (Outcome)	.08	-.03, .18
Unhedged: Outcome	-.99	-1.18, -.82
PropAtt: Outcome	-1.10	-1.27, -.93
Double: Outcome	-.73	-.88, -.57
phi	4.30	4.16, 4.45
zoi	.08	.08, .09
coi	.30	.27, .34

Crucial, here, are the interaction terms (Unhedged: Outcome, PropAtt: Outcome, and Double: Outcome). As we now expect, there is a reputational hit in all the experimental (non-null) conditions. This effect is reliably larger when outcomes are known than when there are merely expectations. But, notably, the interaction term 'Double: Outcome' is the smallest, and is reliably smaller than the term 'PropAtt: Outcome', as shown by non-overlapping 95% credible intervals. In other words, there is less of an increase in reputational hit with Double Hedging than with Propositional Attitude Hedging. These data support the conclusion that there is a different pattern among the experimental (non-null) conditions in Experiment 6 than in Experiment 5.

We tested model fit with posterior predictive checks, as shown in Fig. 8, which reveal over-prediction of ratings in approximately the .15 to .50 range (15 to 50 on the

response scale) and under-prediction of ratings in approximately the .65 to .90 range (65 to 90 on the response scale).

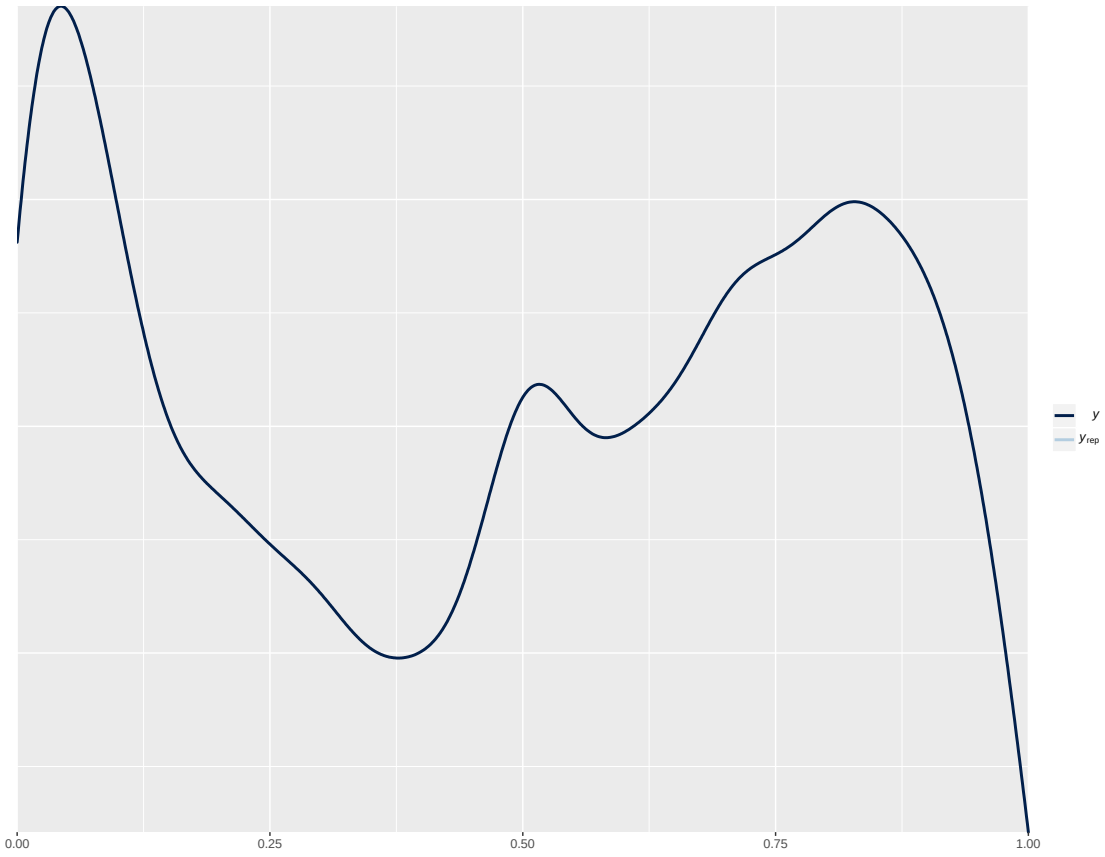

Figure 8. Posterior predictive checks for the model, showing the data (y) against 100 predictions from the model (yrep).

Discussion

Experiment 6 corroborates the findings of the previous experiments. Once outcomes are known, there is still good evidence for the reputational hit and for *Weak Vigilance*: unsurprisingly, there is a considerably larger reputational hit when outcomes are known. Moreover, telling participants the outcome does not reveal evidence to endorse hedging: there is still considerable overlap between Unhedged and Hedged conditions, though with somewhat reduced overlap between Unhedged

and Double Hedged Conditions relative to Experiment 5. There is still, then, no clear evidence for a protective effect of hedging. There is, however, some evidence for an effect of shielding, that is, for a difference between hedging with propositional attitudes and hedging with double hedges. We can see this in the small degree of overlap in the respective 95% HPD intervals in Experiment 6. There also appears to be a somewhat different pattern of results in Experiments 5 and 6. When outcomes are known, there is always an increase in reputational hit over pure expectations, but this increase is smallest in the Double Hedging Condition and is reliably smaller than in the Propositional Attitude Condition. We take this to be weak evidence of a difference between hedging with propositional attitudes and double hedging.

In summary, while introducing outcomes does affect ratings, it does not fundamentally alter our conclusions about the evidence: there remains good evidence for vigilance and no clear evidence for a protective effect of hedging on perceived reliability.

General Discussion

In six experiments, we have explored whether speakers can protect their reputations by hedging claims with evidential language. We have tested hedging against expectation-based and outcome-based updating. We have found no clear evidence for a protective effect of hedging in these experiments, and strong evidence of a reputational hit even when claims are hedged. Our data are unpromising for the view that evidential hedging serves to protect speakers' reputations, whether with a propositional attitude or a double hedge. Successful hedging would have been a way to escape both the confirmation bias seen with expectation-based updating and the

dilemma for game-theoretic accounts of pragmatics. But we find no evidence for successful hedging.

Although our data do not look promising for reputation management with evidential hedges, we do not claim to have falsified it. We have made a number of assumptions to test for such hedging (on auxiliary assumptions and falsification, see, e.g., Earp & Trafimow, 2015; Lakatos, 1978; Meehl, 1990; Trafimow, 2009, 2012). These assumptions have produced a distinctive set of tasks in which we failed to find successful hedging. But varying these assumptions might produce more promising results; we see this paper as part of a broader debate. To motivate future work, we discuss our assumptions, attendant limitations, and future directions. We then consider the possible implications of our data for the literature.

Expectations

There are a number of points to be made about expectations. Firstly, one key assumption was that expectations are a fair test case for reputation management. This assumption is clearest in Experiments 1 to 5, where the materials only manipulated expectations and never specified the outcomes. But the assumption was carried over into Experiment 6: to keep the designs and results as comparable as possible, we fixed the claims in Experiment 6 as low expectedness, and specified the outcome - that the claims proved false.

We have justified the use of expectations on various grounds. We have linked expectations and known outcomes: we see known outcomes as an extreme case of

expectations, and see no reason to sharply distinguish the two from the recipient's perspective (though see the Introduction for a distinction in accuracy). Moreover, expectations are the more general case, as they are implicated in a great many conversations about things whose truth we cannot (immediately) establish. Finally, expectation-based updating seems to pose as much of a challenge to cooperation as outcome-based updating.

But we must acknowledge a limitation with respect to testing McCready's theory (2015). Since McCready's (2015) theory focuses on cases when outcomes are known, more data are needed from experiments that specify outcomes. We found no clear evidence of hedging when recipients learnt that a low expectedness claim proved false. But hedging might prove more effective when claims have medium or high prior probability and yet prove false. For example, if a sports journalist predicts that a horse with an excellent track record will win when it, in fact, loses, then that journalist may suffer reputational damage. But if the journalist hedges the prediction, they may plausibly evade the reputational damage. Future experiments could usefully vary expectations to cover more of the probability scale. We question, though, whether claims that prove false will actually lead to reputational damage at all, hedged or not, when the claim is highly plausible.

Varying expectations brings us to a second key point. Our data suggest that, when speakers make low expectedness claims, they undermine their reputation; and by "low expectedness" we have meant low prior probability. But if speakers are to be highly informative, they must communicate true information that is not just unknown to the hearer but also surprising to them: that has a low prior probability. It might, then, be impossible to be informative without the risk of reputational damage.

What we have not tested, however, is the extent to which that reputational damage scales with the level of probability in question. Our methods could be adapted to vary the expectedness (prior probabilities) alongside knowledge of the outcomes, to explore the full range of the probability scale and identify where, and at what level, reputational hits occur. We suspect that a key issue will be whether speakers provide any supporting information. If speakers make surprising claims without supporting evidence - as characters in our experimental materials did - they may suffer a reputational hit. But perhaps a hit is defused if speakers rely on more complex utterances to make low expectedness claims: claims together with supporting evidence, or with some acknowledgement that the claim is surprising. At the same time, it will be important to examine in future work how reputation recovers when low probability claims ultimately turn out to be true, in particular to what extent they return reputation to a higher level than before.

Hedging

Across all of the experiments, we operationalized hedging in just two ways: with the propositional attitude "suspect" and the double hedge "I might be wrong, but I think...." We have assumed that these hedges fairly represent their respective categories. There is certainly evidence that people consider these hedges weaker than unhedged claims: Collins (2017) reports manipulation checks conducted with Experiments 3 and 4 that show that both hedges were rated as expressing significantly weaker claims than unhedged claims. But there is, of course, a wide range of alternative hedges available. For instance, there are numerous other propositional attitudes of different strengths, such as 'suppose', 'feel', 'intuit', 'reckon', and many more.

It is possible that other hedges might be more effective. But since exploring a wide range hedges comes with a risk of false-positive results (see, e.g., Simmons, Nelson, & Simonsohn, 2011; Wicherts et al., 2016), any such exploration should be underpinned by theory to explain why hedges differ. We suggest that future work should draw more deeply on semantics than we have done in this paper. We have taken a similar path to research on verbal probability expressions, and assumed that the probability (certainty) expression maps onto the probability scale, without articulating a clear semantic (or pragmatic) theory (for discussion, see Collins & Hahn, 2018). A clear theory would be invaluable in predicting which hedges should work and in which contexts. We will suggest some possible ingredients of successful hedging later in this section.

Reliability

Another set of assumptions concerned reliability. Firstly, we implicitly assumed that hedging should work across the range of source reliability, and fixed our sources as reliable across all the experiments to avoid floor effects. This assumption is open to question. There are, in fact, data suggesting that people expect high-expertise (i.e. reliable) sources to be speak with authority or certainty (Longman et al., 2012). Thus, if high-expertise sources communicate uncertainty through hedging, they may actually damage their own reputation rather than protecting it (though see Karmarkar & Tormala, 2010, for counterintuitive effects on persuasion). If hedging made reputations worse in our tasks, we would expect to see hedged conditions rated reliably lower than unhedged conditions, a pattern we did not see. It could nevertheless be informative to include lower-reliability sources.

While our design could have detected a quantitative effect like this, it could not have detected a qualitative shift. We have assumed that people interpret reliability in the same way across conditions: that "reliability" means the same when there are unhedged claims as when there are hedged ones. In Experiments 5 and 6, we included a definition to try to achieve this consistency: a person's "credibility, trustworthiness, or expertise - as whether they tend to say the truth". We consider this definition faithful to the sense of reliability relevant for McCready (2015) and for Bayesian models of testimony. But it does allow some flexibility in interpretation.

It is conceivable that participants focused on different aspects of reliability in the different conditions. For instance, without a hedge, participants may have focused strictly on the source's accuracy; but with a hedge, they may have focused on expertise in a different sense, the source's perceived authoritativeness. After all, the source who hedges may be accurately reporting their mental state, irrespective of the true state of affairs in the world: whether or not valium is a good treatment for severe coughs, the speaker *suspects* that it is. A reputational hit could, then, mean different things in different conditions, with ratings reflecting different aspects of reliability. To detect such qualitative shifts, future experiments could unpack reliability into different scales (as is often done in the persuasion literature; see, e.g., McCroskey & Teven, 1999). More theoretical work would also be needed to identify which aspect of reliability is most important to cooperation.

A second assumption is embodied in the way we conveyed the sources' expertise. We decided against giving participants a sequence of interactions with a particular source from which they could induce the source's reliability. We instead used background knowledge and context to suggest the source's reliability, which

minimized demands on memory and inference. This simplification seems reasonable, as does the assumption that evidential hedging should still apply, but future work could give participants direct experience with a source over multiple trials. Plausibly, the way information is conveyed - by description or experience - could affect how firmly participants fix their initial belief about the source's reliability, hence how prepared they are to shift it based on the low expectedness claim.[14]

Finally, we assumed that reliability could be measured on ratings scales. Such scales are a simple, direct way of operationalizing reliability. These, and similar, scales are widely used in research on persuasion (for reviews of resulting findings, see Briñol & Petty, 2009; Petty & Briñol, 2008), argumentation (e.g. Bhatia & Oaksford, 2015; Corner & Hahn, 2009; Eemeren, Garssen, & Meuffels, 2009; Hahn & Oaksford, 2007), and testimony (e.g. Collins et al., 2018; Harris, Hahn, Madsen, & Hsu, 2016). Other researchers might prefer an alternative dependent measure. For instance, a character's reliability might be inferred from a participant's willingness to cooperate with them in a cooperation game.

Implications

These assumptions suggest the need for a larger research program. But let us assume, for now, that our results stand with the alternative methods and measures mentioned above. What would the implications be? Hedging of some kind is

[14] We thank an anonymous reviewer for this suggestion.

intuitively appealing. We still find it plausible that speakers can somehow protect their reputations with hedging. One possibility is that hedging with propositional attitudes, shield hedges, or double hedging works when there is a clear contrast. In our experiments, even when participants were prompted to attend to the hedges, they had nothing to compare the hedges to. But successful hedging may, in fact, rely on explicit contrasts. Take, for instance, the following exchange:

Bob: Looks like it'll rain tomorrow.

Sandy: Are you sure?

Bob: Well, I suspect it will.

In this context, Bob may be understood as meaning that he is not sure: that he only has sufficient confidence to use 'suspect' (Horn, 1989; Levinson, 2000; Van Der Auwera, 1996; Verstraete, 2005). Hedging might work with such clear contrasts. But McCready's (2015) account is so appealing in large part because hedging provides a way to sustain cooperation. If hedging only works through such pragmatic inferences, it presumably plays less of a role in justifying and maintaining cooperation, since pragmatics is typically taken to be the result of cooperation, not a precondition for it[15]. Nevertheless, one future direction is to develop materials such as the example dialogue above, where a demand for certainty is met with a hedged claim.

Another potential factor can be seen in a parallel with verbal probability expressions: modal adjectives and adverbs, such as 'impossible', 'possibly', 'likely' and 'certain'. These terms feature in standardized lexicons for expressing risk, for

[15] We take McCready (2015) to hold a more nuanced view of her proposed mechanism, one in which the mechanism sits at the boundary between semantics and pragmatics.

instance, the risks of climate change (e.g. Intergovernmental Panel on Climate Change, 2005). One feature of verbal probability expressions is that, independently of the probability they convey, some suggest that the described event will occur and some that it will not. That is, they have directionality. For example, although participants give ‘some possibility’ and ‘quite uncertain’ similar numerical interpretations, significantly more participants prefer an operation described as having ‘some possibility’ of success to one whose success is ‘quite uncertain’ (Teigen & Brun, 1999; for more general discussion of scales of alternatives, see Geurts, 2013).

Thanks to directionality, verbal probability expressions can have subtle effects on reputation. Teigen (1988) presented participants with predictions couched in verbal probability expressions with different directionality: one expert predicted a rise in crude oil prices by saying "It is possible that oil prices will reach $20 in October", another by saying "It is not quite certain that oil prices will reach $20 in October". Participants were then told that oil reached $20. Even though participants thought the first expert had a lower probability in mind than the second, they decided that the first was more right (for discussion, see Teigen & Brun, 1999).

The question naturally arises of whether directionality is essential to hedging. Certainly, the present studies used a propositional attitude with positive directionality. One test for directionality is to think of (or ask participants for) continuations: so, for example, to complete the sentence "It’s possible that X because…" we would add a reason for occurrence; but to complete the sentence "I’m not completely certain that X because…." we would add a reason for non-occurrence. ‘Suspect’ suggests occurrence: "I suspect that X" naturally invites continuations that justify X being the case. The double hedge is more complicated: the sentence "I might be wrong,

because…" invites reasons for wrongness, but the word 'but' seems to override this directionality. Future work could manipulate directionality in hedging with propositional attitudes. However, if such hedging is to play a role in justifying cooperation, an account would also need to show that directionality is a semantic, rather than pragmatic, phenomenon. Geurts (2013) offers hope here, outlining a semantic account of scales of alternatives for various types of expression, including probability statements.

Our data also relate to research on plausible deniability and indirect speech (Lee & Pinker, 2010; Pinker, Nowak, & Lee, 2008). According to this research, a strategic speaker can choose to make indirect speech acts when it is unclear whether a context involves cooperation or conflict. In particular, a strategic speaker can enable a cooperative hearer to act favourably and prevent an uncooperative hearer from acting antagonistically. Imagine, for example, a motorist stopped by a police officer and given a ticket. The motorist wishes to avoid a ticket by offering a bribe. A bribe would be accepted by a dishonest police officer, but would lead to arrest by an honest police officer. One option is to offer a bribe through an indirect speech act, such as "Gee, officer, is there some way we could take care of the ticket here?" (Lee & Pinker, 2010; Pinker, Nowak, & Lee, 2008). This utterance would allow a dishonest officer to recognize and accept the bribe, but prevents an honest officer from having clear evidence for an arrest. There is experimental evidence that indirectness is sensitive to such pay-off structures (Lee & Pinker, 2010).

Strategic indirectness might seem like a case of successful hedging, and hence might seem to conflict with our results. There is, however, no true conflict. Strategic indirectness may well allow plausible deniability, but its effects do not depend on

reputation management. Take the case of the motorist above. While the motorist may well have done enough to avoid being arrested, they are likely to suffer reputational damage. Indeed, such manoeuvring can be transparent and can itself cause reputational damage without undermining its strategic objective.

As we have seen, evidential language might yet allow reputation management but through a pragmatic mechanism: direct comparison with a set of alternative expressions, giving rise to a scalar implicature. But what function might the evidential language fulfil here if not reputation management? The evidential language is presumably not meaningless. We can again draw a comparison with verbal probability expressions. These expressions have uses beyond conveying probabilities (for discussion, see Collins & Hahn, 2018). Bonnefon and Villejoubert (2006) showed that verbal probability expressions can be used to convey bad news tactfully. They had participants read the sentence, "The doctor tells you, you will possibly suffer from insomnia [deafness] soon" and elicited membership functions from participants.[16] Participants understood 'possibly' as indicating higher probabilities with deafness than insomnia. Moreover, some 60% of participants indicated that the doctor was being tactful not uncertain. Bonnefon and Villejoubert argue that different interpretations can result from participants perceiving an utterance as a face-threatening act: that is, as an act that threatens someone's desire for autonomy (their negative face) or their desire for connection with others (their positive face) (Brown

[16] In membership-function studies, participants see a number line representing the probability range [0,1]. The line picks out numbers at even intervals across the range, different studies selecting different intervals. Participants are asked to rate how well each number corresponds to a verbal probability expression. These data allow a membership function to be calculated which can be interpreted as representing the meaning of the expression.

& Levinson, 1987; Holtgraves, 2002). On Bonnefon and Villejoubert's account, verbal probability expressions can serve as face-management devices: they allow a speaker to acknowledge, and lessen the impact of, a face-threatening act.

This politeness account can be extended to the present data. Participants may have understood at least some of the claims as advice: as advice, say, to take a medicine, buy an oven, bet on a horse, and so on. Participants might take this advice as an imposition or threat to their freedom of action: in the terms of Politeness Theory (Brown & Levinson, 1987), as a threat to negative face. Advice is not far removed from requests, offers, compliments, and so on, which are typically taken to threaten negative face (for discussion, see Holtgraves, 2002). Future work could probe this possibility by asking participants why the speaker used the hedges, after Bonnefon and Villejoubert (2006). Although a polite source could seem cooperative, helpful, and so on – hence, in some sense, reliable - they may nevertheless not appear reliable in the sense of being a source of true information. It could be instructive, therefore, to manipulate the social setting and, with it, the plausibility of face-management strategies.

As we have seen, much remains to be explored about reputation management. This work should be part of a broader program of experimental work on the relationship between trust (perceived reliability), cooperation, and pragmatics, addressing the question of to what extent pragmatics can proceed without trust. Much as McCready (2015) seems, to us, to have identified a genuine threat to pragmatics, there are clearly contexts in which pragmatics can proceed with incomplete trust or cooperation. Pragmatics does not stop in arguments or in adversarial situations such as court rooms (Asher & Lascarides, 2013; Goodwin, 2001). And even when a

speaker is, in a sense, as unreliable as possible - is a compulsive liar - their utterances may be informative. In some contexts, the hearer could reverse the content of the message or decrease their belief in what the compulsive liar says. More data are needed on when, and precisely how, distrust undermines pragmatics. Part of the picture will be speakers' attempts to manage their reputations.

Conclusions

Our data suggest that reputation damage persists despite hedging with propositional attitudes and double hedges. This damage occurs both with expectations and known outcomes. The data also cast doubt on the view that evidential hedging serves a strategy for speakers to manage their reputations and protect their reliability as information sources and conversation partners. These data do not support a way to defuse the damaging consequences of expectation-based or outcome-based updating. We have suggested a number of avenues for future work. It remains unclear, though, whether the options we suggest could sustain cooperation, and whether they can offer an escape from the dilemma that faces game-theoretic accounts of pragmatics.

Acknowledgements

Peter Collins was supported by ESRC grant (ES/J500021/1) from the Bloomsbury Doctoral Training Centre, held at Birkbeck, University of London. This grant also funded Experiments 1 to 4. He was also supported by a postdoctoral fellowship at the Munich Center for Mathematical Philosophy, funded by an Alexander von Humboldt Professorship to Professor Stephan Hartmann. Ulrike Hahn was supported by an Annaliese Maier Research Award from the Alexander von

Humboldt Foundation. This grant also funded Experiments 5 and 6. The funders had no involvement in study design; in collection, analysis and interpretation of data; in writing the report; or in the decision to submit the article for publication.

References

Adler, J. (2015). *Epistemological problems of testimony* (Summer 2015; E. N. Zalta, Ed.). Retrieved from https://plato.stanford.edu/archives/sum2015/entries/testimony-episprob/

Asher, N., & Lascarides, A. (2013). Strategic conversation. *Semantics and Pragmatics*, *6*(0), 2-1–62. https://doi.org/10.3765/sp.6.2

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*(2), 527–535. https://doi.org/10.3758/s13428-012-0265-2

Bhatia, J.-S., & Oaksford, M. (2015). Discounting testimony with the argument ad hominem and a Bayesian congruent prior model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1548–1559. https://doi.org/10.1037/xlm0000151

Bonnefon, J.-F., & Villejoubert, G. (2006). Tactful or Doubtful? Expectations of Politeness Explain the Severity Bias in the Interpretation of Probability Phrases. *Psychological Science*, *17*(9), 747–751. https://doi.org/10.1111/j.1467-9280.2006.01776.x

Bovens, L., & Hartmann, S. (2003). *Bayesian Epistemology* [OUP Catalogue]. Retrieved from Oxford University Press website: https://ideas.repec.org/b/oxp/obooks/9780199270408.html

Briñol, P., & Petty, R. E. (2009). Source factors in persuasion: A self-validation approach. *European Review of Social Psychology*, *20*(1), 49–96. https://doi.org/10.1080/10463280802643640

Brown, P., & Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage*. Cambridge, England: Cambridge University Press.

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes*, *36*(3), 391–405. https://doi.org/10.1016/0749-5978(85)90007-X

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Chance, Z., Norton, M. I., Gino, F., & Ariely, D. (2011). Temporal view of the costs and benefits of self-deception. *Proceedings of the National Academy of Sciences*, *108*(Supplement 3), 15655–15659. https://doi.org/10.1073/pnas.1010658108

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

Coady, C. A. J. (1992). *Testimony: A Philosophical Study*. Oxford, England: Oxford University Press.

Collins, P. J. (2017). *Rationality, pragmatics, and sources.* (Unpublished doctoral dissertation, Birkbeck, University of London). Retrieved from http://bbktheses.da.ulcc.ac.uk/284/

Collins, P. J., & Hahn, U. (2018). Communicating and reasoning with verbal probability expressions. In *Psychology of Learning and Motivation* (Vol. 69, pp. 67–105).

Collins, P. J., Hahn, U., von Gerber, Y., & Olsson, E. J. (2015). The bi-directional relationship between source characteristics and message content. In D. C. Noelle, R. Dale, S. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 423–428). Austin, TX: Cognitive Science Society.

Collins, P. J., Hahn, U., von Gerber, Y., & Olsson, E. J. (2018). The Bi-directional Relationship between Source Characteristics and Message Content. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00018

Corner, A., & Hahn, U. (2009). Evaluating science arguments: Evidence, uncertainty, and argument strength. *Journal of Experimental Psychology. Applied*, *15*(3), 199–212. https://doi.org/10.1037/a0016533

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00621

Eemeren, F. H. van, Garssen, B., & Meuffels, B. (2009). *Fallacies and Judgments of Reasonableness: Empirical Research Concerning the Pragma-Dialectical Discussion Rules*. Dordrecht: Springer.

Ganea, P. A., Koenig, M. A., & Millett, K. G. (2011). Changing your mind about things unseen: Toddlers' sensitivity to prior reliability. *Journal of Experimental Child Psychology*, *109*(4), 445–453. https://doi.org/10.1016/j.jecp.2011.02.011

Geurts, B. (2013). Alternatives in Framing and Decision Making. *Mind & Language*, *28*(1), 1–19. https://doi.org/10.1111/mila.12005

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of*

Personality and Social Psychology, *59*(4), 601–613. https://doi.org/10.1037/0022-3514.59.4.601

Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, *65*(2), 221–233. https://doi.org/10.1037/0022-3514.65.2.221

Goodman, N. D., & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Goodwin, J. (2001). The non-cooperative pragmatics of arguing. In E. Nemeth (Ed.), *Pragmatics in 2000: Selected papers from the 7th International Pragmatics Conference.* (Vol. 2, pp. 263–277). Antwerp, Belgium: International Pragmatics Association.

Grice, H. P. (1975). Logic and Conversation. In Davidson, D. & (second) Harman, G. (Eds.), *The logic of grammar* (pp. 64–75). Encino, CA: Dickenson.

Hahn, U., Merdes, C., & von Sydow, M. (2018). How good is your evidence and how would you know? *Topics in Cognitive Science*, *10*(4), 660–678.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, *114*(3), 704–732. https://doi.org/10.1037/0033-295X.114.3.704

Hahn, U., Oaksford, M., & Harris, A. J. L. (2012). Testimony and Argument: A Bayespian Perspective. In F. Zenker (Ed.), *Bayesian Argumentation* (pp. 15–38). New York, NY: Springer.

Harris, A. J. L., Hahn, U., Madsen, J. K., & Hsu, A. S. (2016). The Appeal to Expert Opinion: Quantitative Support for a Bayesian Network Approach. *Cognitive Science*, 1496–1533. https://doi.org/10.1111/cogs.12276

Hasson, U., Simmons, J. P., & Todorov, A. (2005). Believe It or Not On the Possibility of Suspending Belief. *Psychological Science*, *16*(7), 566–571. https://doi.org/10.1111/j.0956-7976.2005.01576.x

Holtgraves, T. M. (2002). *Language As Social Action: Social Psychology and Language Use*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Horn, L. R. (1989). *A natural history of negation*. Chicago, IL: University of Chicago Press.

Intergovernmental Panel on Climate Change. (2005). *Guidance notes for lead authors of the IPCC Fourth Assessment Report on Addressing Uncertainties.* Retrieved from http://www.ipcc-wg2.awi.de/guidancepaper/uncertainty-guidance-note.pdf

Jensen, J. D. (2008). Scientific Uncertainty in News Coverage of Cancer Research: Effects of Hedging on Scientists' and Journalists' Credibility. *Human Communication Research*, *34*(3), 347–369. https://doi.org/10.1111/j.1468-2958.2008.00324.x

Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (In)Accuracy in Perceptions of Groups and Individuals. *Current Directions in Psychological Science*, *24*(6), 490–497. https://doi.org/10.1177/0963721415605257

Karelitz, T. M., & Budescu, D. V. (2004). You Say "Probable" and I Say "Likely": Improving Interpersonal Communication With Verbal Probability Phrases.

Journal of Experimental Psychology: Applied, *10*(1), 25–41. https://doi.org/10.1037/1076-898X.10.1.25

Karmarkar, U. R., & Tormala, Z. L. (2010). Believe Me, I Have No Idea What I'm Talking About: The Effects of Source Certainty on Consumer Involvement and Persuasion. *Journal of Consumer Research*, *36*(6), 1033–1049. https://doi.org/10.1086/648381

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. https://doi.org/10.1037/a0029146

Kuhn, D. (1991). *The Skills of Argument*. Cambridge University Press.

Lakatos, I. (1978). *The methodology of scientific research programmes*. Cambridge, England: Cambridge University Press.

Lassiter, D. (2010). Gradable epistemic modals, probability, and scale structure. *Semantics and Linguistic Theory*, *20*, 197. https://doi.org/10.3765/salt.v20i0.2557

Lassiter, D. (2017). *Graded Modality: Qualitative and Quantitative Perspectives*. Oxford University Press.

Lee, J. J., & Pinker, S. (2010). Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review*, *117*(3), 785–807. https://doi.org/10.1037/a0019688

Lenth, R. (2018). emmeans: Estimated Marginal Means, aka Least-Squares Means. (Version R packages version 1.2.3). Retrieved from https://CRAN.R-project.org/package=emmeans

Levine, G., & Parkinson, S. (1994). *Experimental Methods in Psychology*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, *7*(4), 560–572. https://doi.org/10.1016/0010-0285(75)90023-7

Longman, T., Turner, R. M., King, M., & McCaffery, K. J. (2012). The effects of communicating uncertainty in quantitative health risk estimates. *Patient Education and Counseling*, *89*(2), 252–259. https://doi.org/10.1016/j.pec.2012.07.010

Mandelbaum, E. (2014). Thinking is Believing. *Inquiry*, *57*(1), 55–96. https://doi.org/10.1080/0020174X.2014.858417

Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, *112*(3), 367–380. https://doi.org/10.1016/j.cognition.2009.05.012

Mazar, N., Amir, O., & Ariely, D. (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research*, *45*(6), 633–644. https://doi.org/10.1509/jmkr.45.6.633

McCready, E. (2015). *Reliability in Pragmatics*. Oxford, England: Oxford University Press.

McCroskey, J. C., & Teven, J. J. (1999). Goodwill: A reexamination of the construct and its measurement. *Communication Monographs*, *66*(1), 90–103. https://doi.org/10.1080/03637759909376464

Meehl, P. E. (1990). Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Meredith, M., & Kruschke, J. K. (2013). BEST: Bayesian Estimation Supersedes the t-Test (Version .2.0). Retrieved from http://cran.r-project.org/web/packages/BEST

Mills, C. M., & Keil, F. C. (2005). The Development of Cynicism. *Psychological Science*, *16*(5), 385–390. https://doi.org/10.1111/j.0956-7976.2005.01545.x

Mills, C. M., & Keil, F. C. (2008). Children's developing notions of (im)partiality. *Cognition*, *107*(2), 528–551. https://doi.org/10.1016/j.cognition.2007.11.003

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes Factors for common designs. (Version R package version 0.9.12-2). Retrieved from https://CRAN.R-project.org/package=BayesFactor

Olsson, E. J. (2011). A Simulation Approach to Veritistic Social Epistemology. *Episteme*, *8*(02), 127–143. https://doi.org/10.3366/epi.2011.0012

Olsson, E. J., & Vallinder, A. (2013). Norms of assertion and communication in social networks. *Synthese*, *190*(13), 2557–2571. https://doi.org/10.1007/s11229-013-0313-1

Petty, R. E., & Briñol, P. (2008). Persuasion: From Single to Multiple to Metacognitive Processes. *Perspectives on Psychological Science*, *3*(2), 137–147. https://doi.org/10.1111/j.1745-6916.2008.00071.x

Phillips, N. (2017). Yarrr: A companion to the e-Book "YaRrr!: The Pirate's Guide to R". (Version R package version 0.1.5). Retrieved from https://CRAN.R-project.org/package=yarrr

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences*, *105*(3), 833–838. https://doi.org/10.1073/pnas.0707192105

Poulin-Dubois, D., Brooker, I., & Polonia, A. (2011). Infants prefer to imitate a reliable person. *Infant Behavior and Development*, *34*(2), 303–309. https://doi.org/10.1016/j.infbeh.2011.01.006

Poulin-Dubois, D., & Chow, V. (2009). The effect of a looker's past reliability on infants' reasoning about beliefs. *Developmental Psychology*, *45*(6), 1576–1582. https://doi.org/10.1037/a0016715

R Core Team. (2016). *R: A language and environment for statistical computing.* Retrieved from https://www.R-project.org/

Robinson, E. J., Champion, H., & Mitchell, P. (1999). Children's ability to infer utterance veracity from speaker informedness. *Developmental Psychology*, *35*(2), 535–546. https://doi.org/10.1037/0012-1649.35.2.535

Sabbagh, M. A., & Baldwin, D. A. (2001). Learning words from knowledgeable versus ignorant speakers: Links between preschoolers' theory of mind and semantic development. *Child Development*, *72*(4), 1054–1070.

Schwarzer, G. (2007). meta: An R package for meta-analysis. *R News*, *7*(3), 40–45.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows

Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic Vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Teigen, K. H. (1988). The language of uncertainty. *Acta Psychologica*, *68*(1–3), 27–38. https://doi.org/10.1016/0001-6918(88)90043-1

Teigen, K. H., & Brun, W. (1999). The Directionality of Verbal Probability Expressions: Effects on Decisions, Predictions, and Probabilistic Reasoning. *Organizational Behavior and Human Decision Processes*, *80*(2), 155–190. https://doi.org/10.1006/obhd.1999.2857

Toulmin, S., Rieke, R., & Janik, A. (1979). *An introduction to reasoning*. New York NY: Macmillan.

Trafimow, D. (2009). The Theory of Reasoned Action: A Case Study of Falsification in Psychology. *Theory & Psychology*, *19*(4), 501–518. https://doi.org/10.1177/0959354309336319

Trafimow, D. (2012). The role of auxiliary assumptions for the validity of manipulations and measures. *Theory & Psychology*, *22*(4), 486–498. https://doi.org/10.1177/0959354311429996

Van Der Auwera, J. (1996). Modality: The Three-layered Scalar Square. *Journal of Semantics*, *13*(3), 181–195. https://doi.org/10.1093/jos/13.3.181

Verstraete, J.-C. (2005). Scalar quantity implicatures and the interpretation of modality. *Journal of Pragmatics*, *37*(9), 1401–1418. https://doi.org/10.1016/j.pragma.2005.02.003

Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *The Knowledge Engineering Review*, *10*(01), 43–62. https://doi.org/10.1017/S0269888900007256

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01832